



University of Essex

Department of Economics

Discussion Paper Series

No. 774 October 2015

Nonparametric specification testing via the trinity of tests

Abhimanyu Gupta

Note : The Discussion Papers in this series are prepared by members of the Department of Economics, University of Essex, for private circulation to interested readers. They often represent preliminary reports on work in progress and should therefore be neither quoted nor referred to in published work without the written consent of the author.

Nonparametric specification testing via the trinity of tests

Abhimanyu Gupta^{*†}
Department of Economics
University of Essex, UK

October 22, 2015

Abstract

Tests are developed for inference on a parameter vector whose dimension grows slowly with sample size. The statistics are based on the Lagrange Multiplier, Wald and (pseudo) Likelihood Ratio principles, admit standard normal asymptotic distributions under the null and are straightforward to compute. They are shown to be consistent and possessing non-trivial power against local alternatives. The settings considered include multiple linear regression, panel data models with fixed effects and spatial autoregressions. When a nonparametric regression function is estimated by series, we use our statistics to propose specification tests, and in semiparametric adaptive estimation we provide a test for correct error distribution specification. These tests are nonparametric but handled in practice with parametric techniques. A Monte Carlo study suggests that our tests perform well in finite samples. Two empirical examples use them to test for correct shape of an electricity distribution cost function and linearity and equality of Engel curves.

JEL classifications: C12, C14, C31, C33

Keywords: Increasingly many parameters; central limit theorem; Lagrange Multiplier test; Wald test; Likelihood Ratio test; nonparametric specification testing; series estimation; adaptive estimation; spatial autoregression

^{*}*Email:* a.gupta@essex.ac.uk, *Telephone:* +44 - 120 6872 597

[†]I am grateful for very helpful comments by Svetlana Bryzgalova, Marcus Chambers, Tatiana Komarova, Peter Robinson, Myung Hwan Seo, João Santos Silva, Shruti Sinha, Sorawoot Srisuma and Luke Taylor that improved the paper considerably. I also thank Adonis Yatchew for permission to use the data employed in the empirical examples.

1 Introduction

Many statistical models are parameterized by vectors that increase in dimension with sample size, making the study of asymptotic properties of estimates a nonparametric problem. We are concerned with inference on such growing parameter vectors in these type of models. Our tests statistics have desirable asymptotic properties in such settings and are easy to compute using standard formulae and software. We show that they can be applied to wide variety of problems, including panel data models, spatial autoregressive models, and specification testing in nonparametric regression and adaptive estimation. Throughout the paper we will consider only cases where the parameter space grows slowly with sample size, as opposed to models in which the number of parameters exceeds sample size.

Inference rules are complicated in increasing dimension settings by the fact that while usual Lagrange Multiplier (LM), Wald or pseudo Likelihood Ratio (LR) test statistics for q (fixed) restrictions have an asymptotic χ_q^2 distribution under the null hypothesis (under suitable regularity conditions), if in fact $q \rightarrow \infty$ with sample size this limit distribution no longer holds. However, as in previous literature (cf. de Jong and Bierens (1994), Hong and White (1995)), we are motivated by the well known fact that

$$\frac{\chi_q^2 - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } q \rightarrow \infty,$$

and will justify asymptotic properties for such standardized test statistics.

The literature on multiple regression with increasingly many parameters dates back at least to Huber (1973). Portnoy (1984, 1985) studied more general M -estimates of linear regression with growing dimension, and Andrews (1985) also stressed that frequently the choice of regressors is motivated more by degrees of freedom constraints than actual economic theory, hence the appeal of a theory that permits the number of regressors to be related to sample size. While practitioners can adopt an attitude that permits precise estimation of larger models with more data, arrival at a parsimonious model requires rules of inference. Testing of approximate models also requires such rules, e.g. Berk (1974) considered time series autoregressions with increasing dimension while Robinson (1979) studied models with increasing dimension as approximations to infinite distributed lag systems. More recently, Robinson (2003) examined the problem of estimating the parameters of a single equation in a system of increasingly many equations. A very recent development is interest in spatial autoregressions (SAR) with increasing dimension, treated in Gupta and Robinson (2015a,b), the latter paper permitting a nonlinear regression component of in-

creasing dimension. They point out that SAR models can sometimes give rise to increasing parameter asymptotics quite naturally.

An important context in which models of increasing dimension are estimated is series estimation of nonparametric regression, where a nonparametric regression function is approximated by a growing number of basis functions whose coefficients need to be estimated, see e.g. Andrews (1991), Newey (1997). These models and such an estimation strategy offer an attractive role for inferential rules based on increasingly many restrictions. One use, explored for instance by Eubank and Spiegelman (1990), Wooldridge (1992), Yatchew (1992) and Hong and White (1995) is to test regression function specification, providing an alternative to kernel based nonparametric specification testing, cf. Fan and Li (1996), Zheng (1996) and Lavergne and Vuong (2000). We also provide such tests, but our general results have some differences. We avoid normality of disturbances (Eubank and Spiegelman (1990)), sample splitting (Wooldridge (1992), Yatchew (1992)) and do not impose that the data be generated from an iid process (Hong and White (1995)). Additionally we base our test statistics on the trinity of tests, which can provide as simple, or even simpler, inference than the sums of squares based statistics of Hong and White (1995). Our approach can handle a variety of interesting cases. These include, but are not limited to, tests of significance of nonparametric regressions, tests of linearity against a nonparametric alternative and tests of a partially linear model against a fully nonparametric alternative. Another use that we propose is in testing the unknown specification of the error distribution in the semiparametric series-based adaptive estimation techniques employed by, e.g., Beran (1976), Newey (1988) and Robinson (2005, 2010).

Simplification of nonparametric and semiparametric inference is an important issue in practical work. A recent paper by Akerberg, Chen, and Hahn (2012) notes that the complicated nature of semiparametric methods tends to make practitioners reluctant to use them. They stress cases in which the practitioner can effectively ignore the semiparametric nature of certain econometric problems and simply use formulae derived for parametric cases, thus increasing considerably the appeal of semiparametric methods. Our approach to nonparametric and semiparametric testing is in the same spirit. Effectively the tests of specification boil down to simple inference on coefficients in linear regression models with standard testing principles. We hope that this simplicity adds to the toolkit of ‘parametric-like’ procedures in problems that are not fully parametric.

In Section 2 we introduce the setup as well as two examples. Section 3.1 defines the test statistics and introduces their desirable property that we seek sufficient conditions

for. Section 3.2 contains the asymptotic theory, the conditions of which we illustrate in our examples. Section 4 uses these ideas to propose a simple specification test for use in nonparametric regression, while Section 5 introduces a test for error distribution specification in adaptive estimation. Section 6 contains a Monte Carlo study of finite sample performance, also discussing some implementation issues. In Section 7 we use our tests to determine the shape of a Canadian electricity distribution cost function and to test for linearity and equality of Engel curves from South African data. Section 8 concludes, briefly discussing heteroskedasticity robust versions of our tests. Proofs are in appendices.

2 Setup

We observe a vector $w_i \equiv w_{in}$, $i = 1, \dots, n$, of dimension at least $s + 1$, with s a positive integer. The triangular array setup permits sufficient generality to cover many important cases. The unknown parameter vector $\theta_0 \in \mathbb{R}^s$, is estimated by

$$\tilde{\theta} = \arg \min_{\theta \in \Theta} \mathcal{Q}_n(w_1, \dots, w_n; \theta), \quad (2.1)$$

where $\Theta \subseteq \mathbb{R}^s$. s is regarded as n -dependent, with $s \rightarrow \infty$ as $n \rightarrow \infty$ although explicit reference to this dependence is suppressed for notational convenience, as are the observations w_1, \dots, w_n in the objective function.

Assumption 1. $\mathcal{Q}_n(\theta)$ is convex and twice differentiable in $\theta \in \Theta$, for all sufficiently large n .

In situations where $\tilde{\theta}$ is implicitly defined Θ may be a prescribed compact set. Assumption 1 ensures that $\tilde{\theta}$ exists for sufficiently large n , and allows us to define

$$g_n(\theta) = \frac{\partial \mathcal{Q}_n(\theta)}{\partial \theta}, \quad H_n(\theta) = \frac{\partial^2 \mathcal{Q}_n(\theta)}{\partial \theta \partial \theta'}, \quad (2.2)$$

with primes denoting transposition. Split the parameter vector as $\theta = (\theta'_1, \theta'_2)'$ where θ_1 is $q \times 1$, with $q \leq s$ and $q \rightarrow \infty$ as $n \rightarrow \infty$, and θ_2 is $(s - q) \times 1$. For a generic matrix J define $\|J\| = \{\bar{\eta}(J'J)\}^{\frac{1}{2}}$, where $\bar{\eta}(K)$ (respectively $\underline{\eta}(K)$) denotes the largest (respectively smallest) eigenvalue of a square symmetric matrix K . We are interested in

testing hypotheses of the type

$$H_{0n} \equiv H_0 \quad : \quad \|\theta_{10}\| = 0 \quad (2.3)$$

$$H_{1n} \equiv H_1 \quad : \quad \|\theta_{10}\| \neq 0, \quad (2.4)$$

where $\theta_0 = (\theta'_{10}, \theta'_{20})'$ is the true parameter value. As usual θ can be a transformation of some underlying parameters, so that the above formulation is general enough to cover an increasing number of linear restrictions on model parameters. We also denote the parameter space under the restriction (2.3) as Θ^0 and define the restricted estimate

$$\hat{\theta} = \arg \min_{\theta \in \Theta^0} \mathcal{Q}_n(\theta). \quad (2.5)$$

In the sequel we will assume that θ_0 lies in the interior of Θ^0 , and therefore of Θ .

Example I. Inference in regression models with increasing dimension.

Portnoy (1984, 1985) considers

$$y_n = X_n \beta + v, \quad (2.6)$$

where throughout the paper $y \equiv y_n$ denotes an $n \times 1$ vector of observations, $X \equiv X_n$ an $n \times k$ matrix of exogenous explanatory variables, with $k \rightarrow \infty$ as $n \rightarrow \infty$, and v an unobserved $n \times 1$ disturbance vector with iid elements v_i having mean zero and unit variance. He takes

$$\mathcal{Q}_n^M(\beta) = \frac{1}{n} \sum_{i=1}^n \psi(y_i - x_i' \beta), \quad (2.7)$$

where $x_i \equiv x_{i,n}$ is the i -th column of X' , $y_i \equiv y_{i,n}$ is the i -th column of y and $\psi : \mathbb{R} \rightarrow \mathbb{R}$. Thus $s = k$ and $w_i = (y_i, x_i')'$. The ordinary least squares (OLS) estimate is obtained by taking $\psi(x) = x^2/2$, and we maintain this case in this paper. Panel data models with fixed effects can also be accommodated. Consider a balanced panel with N observations in each of T individual panels, so that $n = NT$. Let y_{tN} be the $N \times 1$ vector of observations on the dependent variable for the t -th panel, where t may correspond to a time period or a more general spatial unit like a school, village or district. Also let $X_{1,tN}$ and $X_{2,N}$ be $N \times k_1$ and $N \times k_2$ matrices of exogenous regressors respectively. $X_{1,tN}$ contains panel-varying regressors while $X_{2,N}$ does not. Consider the model

$$y_{tN} = \iota_N \alpha_t + X_{1,tN} \beta_1 + X_{2,N} \beta_2 + v_{tN}, \quad t = 1, \dots, T \quad (2.8)$$

where v_{tN} is the $N \times 1$ vector of disturbances for each panel, formed of iid components, and ι_N is the $N \times 1$ vector of ones. The α_t , $t = 1, \dots, T$, are scalar fixed effect parameters and β_1 is a $k_1 \times 1$ panel-invariant parameter vector, whereas β_{2t} is a $k_2 \times 1$ parameter vector that can vary with t , so $X_{2,N}$ may be thought of as controlling for ‘quasi’ fixed-effects. Denote $y_n = (y'_{1n}, \dots, y'_{Tn})'$, $X_{1,n} = (X'_{1,1n}, \dots, X'_{1,Tn})'$, $v = (v'_{1n}, \dots, v'_{Tn})'$, $\alpha = (\alpha_1, \dots, \alpha_T)'$ and $\beta_2 = (\beta_{21}, \dots, \beta_{2T})'$. Writing I_T for the $T \times T$ identity matrix we can then stack (2.8) to obtain

$$y_n = (I_T \otimes \iota_N) \alpha + X_{1,n} \beta_1 + (I_T \otimes X_{2,N}) \beta_2 + v, \quad (2.9)$$

which can be written like (2.6) by taking $X_n = (I_T \otimes \iota_N, X_{1,n}, I_T \otimes X_{2,N})$ and $\beta = (\alpha', \beta'_1, \beta'_2)'$, implying $s = k_1 + T(k_2 + 1)$. Again we may dispense with n subscripting for brevity.

Our theory will permit inference on subsets of β of increasing dimension. A question of practical interest is whether the fixed effects α_t in (2.8) are zero, or more generally if they are equal. Thus we may be interested in testing:

$$H_0 : \|\alpha - c\iota_T\| = 0, \quad c \text{ known scalar.} \quad (2.10)$$

Example II. Inference in spatial autoregressive (SAR) models with increasing dimension.

The SAR model was introduced by Cliff and Ord (1973) and has seen heavy use since in modelling of spatial correlation and dependence. For a given set of known weight matrices W_{in} , $i = 1, \dots, p$, whose elements are a measure of economic (not necessarily geographic) distance between units, y_n is modelled as

$$y_n = \sum_{i=1}^p \lambda_i W_{in} y_n + X_n \beta + v, \quad (2.11)$$

The λ_i capture spatial dependence between units. We write $R_n = [W_{1n}y_n, \dots, W_{pn}y_n]$ and $\theta = (\lambda', \beta')'$, where λ has i -th element λ_i , $i = 1, \dots, p$, so $s = p + k$. X_n may also contain spatial lags of regressors, so its columns need not be independent or identically distributed in general. The elements of the W_{in} themselves are usually normalized in some way with a normalization factor that depends on n , indeed some normalization is necessary to identify the λ_i . Thus the triangular array aspect here is not merely a technical generalization but an important feature of the model. Each column of R_n is endogenous, so OLS estimation of θ does not work in general but Lee (2002) showed that for $p = 1$ consistency follows if the elements of W_{in} are $\mathcal{O}(h^{-1})$ for some $h \rightarrow \infty$ such that $h = o(n)$,

and asymptotic normality and efficiency if also $n^{\frac{1}{2}} = o(h)$. The asymptotic properties of the instrumental variables (IV) estimate were justified by Kelejian and Prucha (1998), while Lee (2004) presented a number of results for Gaussian (although Gaussianity is nowhere assumed) pseudo maximum likelihood estimates (PMLE). Gupta and Robinson (2015a,b) have introduced an increasing dimension ($p, k \rightarrow \infty$) version of higher-order models such as (2.11) that contain more than one spatial lag of y_n . They consider OLS, IV and PML estimates of θ . The latter works also when $\beta = 0$, unlike the first two, which require the presence of at least one non intercept regressor. Let Z_n be an $n \times r$ matrix of instruments, $r \geq p$, write $\mathcal{P}(J) = J(J'J)^{-1}J'$ for a matrix J with full column rank and define $\hat{\theta}_{SAR}^{IV}$ and $\hat{\theta}_{SAR}^{OLS}$ as the θ minimizing

$$\mathcal{Q}_n^{SAR,IV}(\theta) = \frac{1}{2n} (y_n - [R_n, X_n] \theta)' \mathcal{P}([Z_n, X_n]) (y_n - [R_n, X_n] \theta), \quad (2.12)$$

$$\mathcal{Q}_n^{SAR,OLS}(\theta) = \frac{1}{2n} (y_n - [R_n, X_n] \theta)' (y_n - [R_n, X_n] \theta), \quad (2.13)$$

respectively. Denoting $S_n(\lambda) = I_n - \sum_{i=1}^p \lambda_i W_{in}$, the PMLE is based on the Gaussian likelihood

$$\log(2\pi) - 2n^{-1} \log |S_n(\lambda)| + n^{-1} \|S_n(\lambda) y_n - X_n \beta\|^2. \quad (2.14)$$

For given λ , (2.14) is minimised with respect to β by $\bar{\beta}(\lambda) = (X_n' X_n)^{-1} X_n' S_n(\lambda) y_n$. We define $\hat{\lambda}_{SAR}^{PML} = \arg \min_{\lambda \in \Lambda} \mathcal{Q}_n^{SAR,PML}(\lambda)$, with Λ a compact set in \mathbb{R}^p , where

$$\mathcal{Q}_n^{SAR,PML}(\lambda) = n^{-1} \log |S_n^{-1}(\lambda) S_n^{-1'}(\lambda)| + n^{-1} y_n' S_n'(\lambda) (I_n - \mathcal{P}(X_n)) S_n(\lambda) y_n, \quad (2.15)$$

The PMLE of β_0 is defined as $\bar{\beta}(\hat{\lambda}) \equiv \hat{\beta}_{SAR}^{PML}$.

Gupta and Robinson (2015a,b) stress cases motivated by Case (1991, 1992) in which the W_{in} have a ‘single non-zero diagonal block’ structure. In such cases it is explicitly assumed that there are no spatial effects between units not in the same block, and it is reasonable to expect that the λ_i vary across blocks. However there can be reasons (e.g. geographic or demographic) for practitioners to suspect that some of the λ_i may be equal. Of particular interest is the case where all the λ_i are equal, implying a simpler model in which $p = 1$, and a model of fixed dimension if in fact k is fixed. This kind of test can be captured in the null hypothesis

$$H_0 : \|\lambda - c\iota_p\| = 0, \quad c \text{ known scalar.} \quad (2.16)$$

3 Trinity motivated tests

3.1 The statistics and a desirable property

For any function $f(\cdot)$ we will write $f(\check{\theta}) \equiv \check{f}$ and $f(\theta_0) \equiv f$, where $\check{\theta}$ is a generic estimate of θ , following this convention throughout the paper. Define the standardized LM, Wald and LR test statistics

$$\mathbf{LM}_n = \frac{n\hat{g}'_{1,n}\hat{H}_n^{11}\hat{g}_{1,n} - q}{2^{\frac{1}{2}}q^{\frac{1}{2}}}, \quad (3.1)$$

$$\mathbf{W}_n = \frac{n\check{\theta}'_1 \left(\tilde{H}_n^{11}\right)^{-1} \check{\theta}_1 - q}{2^{\frac{1}{2}}q^{\frac{1}{2}}}, \quad (3.2)$$

$$\mathbf{LR}_n = \frac{2n \left(\tilde{\mathcal{Q}}_n - \hat{\mathcal{Q}}_n\right) - q}{2^{\frac{1}{2}}q^{\frac{1}{2}}}. \quad (3.3)$$

with $\hat{g}_n = (\hat{g}'_{1,n}, \hat{g}'_{2,n})'$ and

$$\hat{H}_n^{-1} = \begin{bmatrix} \hat{H}_n^{11} & \hat{H}_n^{12} \\ \hat{H}_n^{21} & \hat{H}_n^{22} \end{bmatrix},$$

where $\hat{g}_{1,n}$ is $q \times 1$, $\hat{g}_{2,n}$ is $(s - q) \times 1$, \hat{H}_n^{11} is $q \times q$, \hat{H}_n^{12} is $q \times (s - q)$ and \hat{H}_n^{22} is $(s - q) \times (s - q)$. This convention for partitioning is adopted throughout so we will denote

$$J_n = \begin{bmatrix} J_{11,n} & J_{12,n} \\ J_{21,n} & J_{22,n} \end{bmatrix}, \quad J_n^{-1} = \begin{bmatrix} J_n^{11} & J_n^{12} \\ J_n^{21} & J_n^{22} \end{bmatrix}$$

for a generic nonsingular $s \times s$ matrix J_n . LM tests have the favourable feature of requiring estimation of the model only under the null hypothesis, which yields a more parsimonious null model always. In the increasing parameter context sometimes even a finite dimensional null model may be implied, cf. Example II. As usual the Wald statistic is based on the unrestricted estimates alone, while the LR statistic is based on both unrestricted and restricted estimates.

We seek to provide sufficient conditions for three desirable features of the test statistics defined above, encapsulated in the following definition.

Definition A sequence of random variables (test statistics) \mathfrak{A}_n is said to have Property \mathcal{C} if

C.1. Under H_0

$$\mathfrak{A}_n \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \rightarrow \infty.$$

C.2. Under H_1 and $q^{\frac{1}{4}}/n^{\frac{1}{2}} \rightarrow 0$ as $n \rightarrow \infty$, for any $\varrho > 0$,

$$P(|\mathfrak{A}_n| > \varrho) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

C.3. Define $H_{\ell 1} \equiv H_{\ell 1, n} : \theta_{\ell, 10} = \theta_{\ell, 10, n} = \delta_n q^{\frac{1}{4}} / (n \delta_n' \Gamma_n \delta_n)^{\frac{1}{2}}$, where Γ_n is a constant symmetric $q \times q$ matrix with $\underline{\lim}_{n \rightarrow \infty} \underline{\eta}(\Gamma_n) > 0$, $\overline{\lim}_{n \rightarrow \infty} \overline{\eta}(\Gamma_n) < \infty$ and δ_n a constant $q \times 1$ vector such that $\|\delta_n\| = 1$. Then, under the sequence of local alternatives $H_{\ell 1}$ and $q^{\frac{1}{4}}/n^{\frac{1}{2}} \rightarrow 0$ as $n \rightarrow \infty$,

$$\mathfrak{A}_n \xrightarrow{d} \mathcal{N}(2^{-\frac{1}{2}}, 1), \text{ as } n \rightarrow \infty.$$

The conditions specify the asymptotic distribution of \mathfrak{A}_n under H_0 , the consistency of the test based on \mathfrak{A}_n as test statistic, and the fact that such a test can detect local alternatives at $\mathcal{O}\left(q^{\frac{1}{4}}/n^{\frac{1}{2}}\right)$ rate.

3.2 Asymptotic theory

In this section we describe the asymptotic behaviour of the test statistics when g_n asymptotically differs negligibly from a linear function of an $n \times 1$ unobservable disturbance vector ϵ with elements ϵ_i . Let C denote a generic constant, arbitrarily large but independent of n .

Assumption 2. The elements of ϵ are independent with zero mean, unit variance, finite third and fourth moments μ_3 and μ_4 , and satisfy $\max_{1 \leq i \leq n} \mathbb{E}|\epsilon_i|^{4+\chi} \leq C$, for some $\chi > 0$.

Imposing unit variance simplifies our notation but is not restrictive as all results hold with independent homoskedastic disturbances, the latter simply adding another layer of derivations in the proofs and thus avoided here, but discussed in examples. Heteroskedasticity robustness is discussed in Section 8. Introduce an $n \times s$ matrix M_n and the $s \times s$ constant and symmetric matrix $L_n = \mathbb{E}(n^{-1} M_n' M_n)$ satisfying the following assumption:

Assumption 3. The elements of M_n are independent of ϵ , and there exists a sequence $m \equiv m_n$, divergent or bounded, such that their second moments are uniformly $\mathcal{O}(m^2)$. The eigenvalues of L_n are such that $\overline{\lim}_{n \rightarrow \infty} \overline{\eta}(L_n) < \infty$ and $\underline{\lim}_{n \rightarrow \infty} \underline{\eta}(L_n) > 0$.

The condition on the elements of M_n implies that these are uniformly $\mathcal{O}_p(m)$ and the rows of M_n have uniformly $\mathcal{O}_p\left(s^{\frac{1}{2}}m\right)$ norm. The restrictions on the eigenvalues of L_n are asymptotic ‘no multicollinearity’ and boundedness conditions of the type familiar from regression with increasing dimension. For generic matrices, vectors or scalars J_n , K_n we denote $\Delta_K^J = J_n - K_n$ and for any symmetric matrix J_n partitioned in the usual way, denote $\mathcal{X}_n^J = [I_q, -J_{12,n}J_{22,n}^{-1}]$.

Theorem 3.1. *Suppose that Assumptions 1, 2 and 3 hold. Suppose also that*

$$\begin{aligned} & \frac{1}{q^{\frac{1}{2}}} \left\| \Delta_{n^{-1}M'\epsilon}^g \right\| \left(n \left\| \Delta_{n^{-1}M'\epsilon}^g \right\| + \|M_n'\epsilon\| \right) + \left\| \Delta_H^{H*} \right\| + \left\| \Delta_{n^{-1}M'M}^H \right\| + \left\| \Delta_L^{n^{-1}M'M} \right\| \\ & + \frac{1}{q^{\frac{1}{2}}n} \|M_n'\epsilon\|^2 \left(\left\| \Delta_H^{H*} \right\| + \left\| \Delta_{n^{-1}M'M}^H \right\| + \left\| \Delta_L^{n^{-1}M'M} \right\| \right) = o_p(1), \text{ as } n \rightarrow \infty, \end{aligned} \quad (3.4)$$

for any θ^* satisfying $\left\| \Delta_{\theta_0}^{\theta^*} \right\| \leq \left\| \Delta_{\theta_0}^{\hat{\theta}} \right\|$. Then

$$\mathbf{LM}_n - \frac{\epsilon' \mathcal{M}_n \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = o_p(1), \text{ as } n \rightarrow \infty,$$

where $\mathcal{M}_n = n^{-1}M_n \mathcal{X}_n^{n^{-1}M'M} (n^{-1}M_n' M_n)^{11} \mathcal{X}_n^{n^{-1}M'M} M_n'$.

Condition (3.4) looks intimidating but is essentially about some weak laws of large numbers holding. We discuss it for special cases in Remarks 2 and 3 below. For a generic $s \times s$ matrix J_n , define Σ_n^J as the $s \times s$ matrix with bottom-right $(s - q) \times (s - q)$ block $J_{22,n}^{-1}$ and all other entries zero.

Theorem 3.2. *Suppose that Assumptions 1, 2 and 3 hold. Suppose also that (3.4) holds for any θ^* satisfying $\left\| \Delta_{\theta_0}^{\theta^*} \right\| \leq \left\| \Delta_{\theta_0}^{\hat{\theta}} \right\|$ or $\left\| \Delta_{\theta_0}^{\theta^*} \right\| \leq \left\| \Delta_{\theta_0}^{\hat{\theta}} \right\|$. Then*

$$\mathbf{W}_n - \frac{n^{-1}\epsilon' M_n \left[(n^{-1}M_n' M_n)^{-1} - \Sigma_n^{n^{-1}M'M} \right] M_n' \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = o_p(1), \text{ as } n \rightarrow \infty, \quad (3.5)$$

$$\mathbf{LR}_n - \frac{n^{-1}\epsilon' M_n \left[(n^{-1}M_n' M_n)^{-1} - \Sigma_n^{n^{-1}M'M} \right] M_n' \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = o_p(1), \text{ as } n \rightarrow \infty. \quad (3.6)$$

In Theorems 3.1 and 3.2 we showed that the trio of test statistics may be approximated by a quadratic form in ϵ . The next two theorems state the asymptotic distribution of these quadratic forms under suitable conditions.

Theorem 3.3. *Suppose that Assumptions 1, 2 and 3 hold. Suppose also that*

$$\frac{1}{q^{1+\frac{\chi}{4}}} \left(\frac{\left(s^{\frac{1}{2}}m\right)^{2+\frac{\chi}{2}}}{n^{\frac{\chi}{4}}} + \frac{\left(s^{\frac{1}{2}}m\right)^{4+\chi}}{n^{1+\frac{\chi}{2}}} \right) \longrightarrow 0, \text{ as } n \rightarrow \infty. \quad (3.7)$$

Then

$$\frac{\epsilon' \mathcal{M}_n \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \rightarrow \infty.$$

The proof is in Appendix A and employs a martingale CLT of Scott (1973) as opposed to the U -statistic CLTs used in earlier literature, cf. e.g. de Jong and Bierens (1994), Hong and White (1995).

Theorem 3.4. *Suppose that Assumptions 1, 2 and 3 hold. Suppose also that (3.7) holds. Then*

$$\frac{n^{-1} \epsilon' M_n \left[(n^{-1} M'_n M_n)^{-1} - \Sigma_n^{n-1} M' M \right] M'_n \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \rightarrow \infty.$$

Remark 1. In all of the theorems above, if $g_n(\theta)$ is linear in θ , the part of the rate conditions (3.4), and (3.7) relating to norm consistent θ^* is not needed. Indeed, in this case there will be a closed form for $\hat{\theta}$ and $\tilde{\theta}$, and the second derivative will not depend on θ .

The following theorem records sufficient conditions for Property \mathcal{C} to hold.

Theorem 3.5. (i) *Under the conditions of Theorems 3.1 and 3.3, \mathbf{LM}_n has Property \mathcal{C} .*

(ii) *Under the conditions of Theorems 3.2 and 3.4, \mathbf{W}_n and \mathbf{LR}_n have Property \mathcal{C} .*

The appropriate choices of Γ_n in $\mathcal{C}.3$ are found in the proof, while δ_n would typically be the vector with unity in the position corresponding to the direction of departures from H_0 . In the remarks that follow we focus only on rate conditions, assuming the identification conditions to hold.

Remark 2. In Example I $g_n = n^{-1} M'_n v$, $M_n = X_n$ and $H_n = n^{-1} X'_n X_n$. If the x_i are iid with finite fourth moment m is a bounded sequence, $\Xi_n = L_n = \mathbb{E}(x_i x'_i)$, $\|M' v\| = \mathcal{O}_p\left((kn)^{\frac{1}{2}}\right)$ and $\|\Delta_L^H\| = \mathcal{O}_p\left(k/n^{\frac{1}{2}}\right)$, so (3.4) holds if $k^2/q^{\frac{1}{2}} n^{\frac{1}{2}} \rightarrow 0$ as $n \rightarrow \infty$. If v_i have constant variance σ^2 , this can be estimated by $\check{\sigma}^2 = \check{\sigma}_n^2 = n^{-1} (y_n - X_n \check{\beta})' (y_n - X_n \check{\beta})$ (where $\check{\cdot}$ can be either \sim or \wedge) and it then standardly follows that $\check{\sigma}^2 - \sigma^2 = \mathcal{O}_p(k/n)$ at worst.

Remark 3. In Example II, writing $R_n = A_n + B_n$ with $A_n = [G_{1n}X_n\beta_0, \dots, G_{pn}X_n\beta_0]$, $B_n = [G_{1n}v, \dots, G_{pn}v]$ and $G_{in} = W_{in}S_n^{-1}$, $i = 1, \dots, p$, we have $g_n = -n^{-1} [R_n, X_n]' v$, $M_n = -[A_n, X_n]$, $H_n = -n^{-1} [R_n, X_n]' [R_n, X_n]$ for OLS and $g_n = -n^{-1} [R_n, X_n]' \mathcal{P}([Z_n, X_n]) v$, $M_n = -\mathcal{P}([Z_n, X_n]) [A_n, X_n]$, $H_n = -n^{-1} [R_n, X_n]' \mathcal{P}([Z_n, X_n]) [R_n, X_n]$ for IV. For PMLE, if $h \rightarrow \infty$ the same analysis as the OLS holds.

Under the conditions of Gupta and Robinson (2015a), we have for OLS: $\|\Delta_{n^{-1}M'v}^g\| = \mathcal{O}_p(p^{1/2}/h)$, $\|M'_n v\| = \mathcal{O}_p(k(pn)^{1/2})$, $\|\Delta_{n^{-1}M'M}^H\| = \mathcal{O}_p(\max\{p/h, p^{1/2}k(p+k)^{1/2}/n^{1/2}\})$, and for IV: $\|\Delta_{n^{-1}M'v}^g\| = \mathcal{O}_p(p^{1/2}(r+k)/n)$, $\|\Delta_{n^{-1}M'M}^H\| = \mathcal{O}_p(p^{1/2}(r+k)^{1/2}/n^{1/2})$ and $\|M'_n v\| = \mathcal{O}_p((r+k)^{1/2}n^{1/2})$. They also assume $\max\{p^{1/2}k, n^{1/2}p^{1/2}h^{-1}\} \|\Delta_L^{n^{-1}M'M}\| = o_p(1)$ for OLS and $(r+k)^{1/2} \|\Delta_L^{n^{-1}M'M}\| = o_p(1)$ for IV. So (3.4) holds under the following conditions as $n \rightarrow \infty$

$$\text{OLS: } \frac{1}{q^{1/2}} \left(\frac{pkn^{1/2}}{h} + \frac{p^{3/2}k^3(p+k)^{1/2}}{n^{1/2}} \right) + \frac{p}{h} + \frac{p^{1/2}k(p+k)^{1/2}}{n^{1/2}} \longrightarrow 0, \quad (3.8)$$

$$\text{IV: } \frac{p^{1/2}(r+k)^{3/2}}{q^{1/2}n^{1/2}} + \frac{p^{1/2}(r+k)^{1/2}}{n^{1/2}} \longrightarrow 0. \quad (3.9)$$

As far as relaxation to $\mathbb{E}(v_i^2) = \sigma^2$ is concerned, taking

$$\begin{aligned} \hat{\sigma}_{SAR}^{2,OLS} &= n^{-1} \left(y_n - [R_n, X_n] \hat{\theta}_{SAR}^{OLS} \right)' \left(y_n - [R_n, X_n] \hat{\theta}_{SAR}^{OLS} \right), \\ \hat{\sigma}_{SAR}^{2,IV} &= n^{-1} \left(y_n - [R_n, X_n] \hat{\theta}_{SAR}^{IV} \right)' \left(y_n - [R_n, X_n] \hat{\theta}_{SAR}^{IV} \right), \end{aligned}$$

By Theorems 3.2, 4.2 of Gupta and Robinson (2015a), $\hat{\sigma}_{SAR}^{2,IV} - \sigma^2 = \mathcal{O}_p((p+k)(r+k)/n)$ and $\hat{\sigma}_{SAR}^{2,OLS} - \sigma^2 = \mathcal{O}_p(\max\{pk^2(p+k)/n, p/h\})$

4 Nonparametric regression specification testing with Property \mathcal{C}

Sometimes it is not reasonable to assume a particular parametric form for the regression function, leading to consideration of

$$y_i = d(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (4.1)$$

with y_i observable, x_i an $k \times 1$ vector of exogenous explanatory variables and $d(\cdot)$ an unknown real-valued function on the support \mathfrak{X} of x_i . On the other hand for multiple regression models such as (2.6), significance of the regression function can be tested by the null $H_0 : \|\beta\| = 0$ and various tests are available for functional form. We propose a specification test based on series estimation. The latter approximates $d(x)$ by $\alpha'_J P_J(x)$, with $P_J(x) = (p_{1J}(x), \dots, p_{JJ}(x))'$ and $\alpha_J = (\alpha_{1J}, \dots, \alpha_{JJ})'$ being $J \times 1$ vectors of basis functions and unknown parameters respectively, and $J \rightarrow \infty$ slowly with n . Given an estimate $\hat{\alpha}_J$, we define a series estimate of $d(x)$ as

$$\hat{d}(x) = \hat{\alpha}'_J P_J(x). \quad (4.2)$$

Andrews (1991) establishes a set of asymptotic normality results for more general settings including functionals of $d(\cdot)$, while uniform convergence rates of $\hat{d}(x)$ to $d(x)$ are derived by Newey (1997) in settings where the data are iid. Our interest is in testing null hypotheses on $d(x)$, which we will test by way of the ‘approximate’ null

$$H_0^{app} : \|\alpha_{t(J)} - c_{t(J)}\| = 0, \quad (4.3)$$

with $\alpha_{t(J)}$ a $t(J) \times 1$ subvector of α_J and $c_{t(J)}$ some known constant $t(J) \times 1$ vector, with $t(\cdot)$ an increasing integer valued (if the image is not an integer we choose the integer part of it) function. The alternative hypothesis will always be the negation of the null being tested.

This setup falls into the framework considered in Section 3.2. Indeed, $\hat{\alpha}_J$ can be formed by least squares regression of $y = (y_1, \dots, y_n)'$ on $P = P_n(x_1, \dots, x_n) = [P_J(x_1), \dots, P_J(x_n)]'$, i.e. $\hat{\alpha}_J = (P'P)^{-1} P'y$, for sufficiently large n if

$$\lim_{n \rightarrow \infty} \underline{\eta} \left(\mathbb{E} \left(P_J(x_i) P_J(x_i)' \right) \right) > 0. \quad (4.4)$$

Thus take $\mathcal{Q}_n(\alpha_J)$ as in (2.7) with $\psi(x) = x^2/2$, implying that $g_n(\alpha_J) = n^{-1} P' (P\alpha_J - y)$ and $H_n(\alpha_J) = n^{-1} P' P$, which does not depend on α_J . Assuming the existence of an $J \times 1$ constant vector α_{0J} and scalar $\gamma > 0$ such that

$$\sup_{x \in \mathfrak{X}} |d(x) - \alpha'_{0J} P_J(x)| = \mathcal{O}_p(J^{-\gamma}), \text{ as } J \rightarrow \infty, \quad (4.5)$$

cf. Newey (1997), and substituting (4.1) gives $g_n = n^{-1} P' (P\alpha_{0J} - D(x) - v)$, with $D(x) =$

$(d(x_1), \dots, d(x_n))'$, whence

$$\|\Delta_{-2n^{-1}P'v}^g\| \leq n^{-1} \|P\| \|P\alpha_{0J} - D(x)\| = \mathcal{O}_p(J^{-\gamma}). \quad (4.6)$$

Under regularity conditions as in Newey (1997), Assumption 3 is satisfied with m a constant sequence.

One case of interest is testing the significance of the nonparametric function, in which case take

$$H_0^{sig} : d(x) = 0, \quad (4.7)$$

$c_J = 0$ (so $\mathbf{t}(J) = J$) in (4.3). If the series functions are polynomials, we can use (4.3) as a more general specification test. To define these, let $\rho = (\rho_1, \dots, \rho_k)'$ be a multi-index with nonnegative entries, z a $k \times 1$ real vector and denote $z^\rho = \prod_{j=1}^k z_j^{\rho_j}$. For a sequence $\{\rho(l)\}_{l=1}^\infty$ of distinct such vectors take $p_{jJ}(x) = x^{\rho(j)}$. For instance, we may test for a linear regression

$$H_0^{lin} : d(x) = \alpha'x, \quad (4.8)$$

by taking $c_{\mathbf{t}(J)}$ to have zeros corresponding to indices j for which $\rho(j)$ is not of the form whose only nonzero entry is unity, implying $\mathbf{t}(J) = J - k$. In practice we must take $J > k$, which is satisfied in the theory as k is fixed. R -th order polynomial regression can also be tested, taking $k = 1$ for simplicity in notation:

$$H_0^{poly} : d(x) = \sum_{i=1}^R \alpha_i x^i, \quad (4.9)$$

and $c_{\mathbf{t}(J)}$ to have zeros corresponding to indices j for which $\rho(j)$ has elements bigger than R , so $\mathbf{t}(J) = J - R$. For general regression of order R with k explanatory variables the multinomial formula gives $\mathbf{t}(J) = J - (k + R - 1)! / (k - 1)! R!$. Other types of basis functions $p_{jJ}(x)$ can yield specification tests against more functional parametric forms, indeed there is no dearth of options for the practitioner. To obtain tests of significance of a subset of regressors in x_i as in Lavergne and Vuong (2000), we can test if all the α_{jJ} corresponding to this subset are zero. Suppose $k = 2$ and we are interested in testing if x_{2i} is a significant regressor. With a polynomial basis we would test if each α_{jJ} that is a coefficient of a product involving x_{2i} is zero.

An important alternative to (4.1) is the partly parametric model (cf. Robinson (1988),

Fan and Li (1996)) where

$$y_i = x'_{1i}\beta + d(x_{2i}) + v_i, \quad i = 1, \dots, n, \quad (4.10)$$

with x_{1i} and x_{2i} subvectors of x_i . Our approach permits a simple way to test against this alternative. The practitioner only has to follow the method described in the paragraph above to test for linearity of the regressors in x_{2i} . But we can offer something more if the linearity part of (4.10) is believed to hold. Our method gives a straightforward way to test for the specification of $d(\cdot)$, and this can be done in even more general cases where $x'_{1i}\beta$ is replaced by a parametric nonlinear function, cf. Andrews (1994).

Use of any of \mathbf{LM}_n , \mathbf{W}_n or \mathbf{LR}_n under H_0^{app} provides a test for the question of interest with Property \mathcal{C} . This has some advantages compared to competing kernel based nonparametric specification tests. There is no need to choose a kernel (although J needs to be chosen in practice), and the statistics are extremely simple to compute.

Proposition 4.1. *Let (y_i, x_i) be iid with*

$$\mathbb{E}(p_{jJ}^4(x)) \leq C, \quad x \in \mathfrak{X}, \quad j = 1, \dots, J, \quad (4.11)$$

Assumption 2, (4.4) and (4.5) hold, $\sup_{x \in \mathfrak{X}} \|P_J(x)\| = \mathcal{O}_p(\zeta(J))$ for some function $\zeta(\cdot)$, and J be chosen as function of n satisfying (3.7) with m constant, $s = J$, $q = \mathfrak{t}(J)$ and

$$\frac{\zeta^4(J)J^2}{n} + \frac{n}{J^{2\gamma}} + \frac{n}{\mathfrak{t}^{\frac{1}{2}}(J)J^{2\gamma-1}} \longrightarrow 0, \quad \text{as } n \rightarrow \infty. \quad (4.12)$$

Then \mathbf{LM}_n , \mathbf{W}_n and \mathbf{LR}_n have Property \mathcal{C} .

When \mathfrak{X} is compact and connected and the support of x_i has a pdf that is bounded away from zero, Newey (1997) derives $\zeta(J) = J$ for the power series basis and $\zeta(J) = J^{\frac{1}{2}}$ for a spline basis, the latter additionally assuming that \mathfrak{X} is known. The value of γ depends on features of $d(x)$ such as smoothness. For power series and spline bases, $\gamma = r/k$, where r is the number of continuous derivatives of $d(x)$ on \mathfrak{X} , cf. Lorentz (1986). Thus (4.12) can hold if $d(x)$ is smooth enough, for given k and choice of basis.

An outstanding practical issue in series estimation is choice of J . Robinson (2005, 2010) discusses this, stating that asymptotic theory provides little guidance as it provides upper but not lower bounds. He points out that these upper bounds suggest very slow increase of J and numerical experiments show that small integer choices of J work well in practice.

Our results can also be used to provide consistent tests with local power for determining J in practice, by simple tests of significance.

The idea applies to more general models. Lee and Robinson (2013) consider series estimation in the presence of cross-sectional dependence, and we can easily accommodate this with (4.12) amended to the correct rates. Indeed, following that paper let x_i be identically distributed with pdf $f(x)$ and the joint pdf of x_i, x_j be $f_{ij}(x, y)$, $x, y \in \mathfrak{X}$. Introduce the bivariate dependence measure

$$\Upsilon_n = \sum_{\substack{i,j=1 \\ i \neq j}}^n \int_{\mathfrak{X} \times \mathfrak{X}} |f_{ij}(x, y) - f(x)f(y)| dx dy, \quad (4.13)$$

and replace (4.12) with

$$\frac{\zeta^6(J)J^4}{n} + \frac{n\zeta^2(J)}{J^{2\gamma-1}} + \frac{\zeta^3(J)J^2\Upsilon_n^{\frac{1}{2}}}{n} \longrightarrow 0, \text{ as } n \rightarrow \infty, \quad (4.14)$$

to obtain tests with Property \mathcal{C} , assuming Υ_n affords a non-null choice set of γ . The latter is obtained from Assumption B.3 in Lee and Robinson (2013), noting that we have taken iid disturbances as opposed to their more general linear process specification. Indeed specification testing can undoubtedly extend to models such as the SAR in (2.11), where the linear regression component is replaced by a nonparametric function which is then estimated by series and tested for linearity as in the previous paragraph. To the best of our knowledge the literature has not yet considered series estimation of the regression function in this setting, with kernel estimation seemingly the preferred tactic (cf. Su and Jin (2010), Jenish (2014)). The approach can also be used to test the specification of any of the fixed number of unknown varying coefficient functions $\vartheta_j(\cdot)$, $j = 1, \dots, r_1$, in the model

$$y_i = x'_{1i}\beta + x'_{2i}\vartheta(x_{3i}) + v_i, \quad i = 1, \dots, n,$$

estimated using the series method by Ahmad, Leelahanon, and Li (2005), where x_{2i} and x_{3i} are $r_1 \times 1$ and $r_2 \times 1$ (r_1, r_2 fixed) vectors of exogenous explanatory variables respectively and $\vartheta(x_{3i}) = (\vartheta_1(x_{3i}), \dots, \vartheta_{r_1}(x_{3i}))'$.

5 Error distribution specification in adaptive estimation

In the adaptive estimation methodology of Newey (1988), which improves upon a treatment of Beran (1976), (2.6) is considered with unknown nonparametric density for a representative element v_i of v . The aim is to obtain efficient estimates of β by means of a Newton-type step that ‘adapts’ to the unknown error density using a series approximation, starting from an inefficient $n^{\frac{1}{2}}$ -consistent initial estimate (such as OLS). It turns out that the object that must be nonparametrically estimated is not the density $f(t)$, but the score function $\varsigma(t) = -f_t(t)/f(t)$, where the t subscript denotes partial derivative with respect to t . There are advantages to using series estimation for this, cf. Robinson (2010) p. 7 for details. To maintain simplicity we take X to consist of uniformly bounded constants (as in Robinson (2010)). Let $\phi_\ell(t)$, $\ell = 1, 2, \dots$, be a sequence of smooth functions and let $J \geq 1$ be some user-chosen integer that increases slowly with n . Define $\phi^{(J)}(t) = (\phi_1(t), \dots, \phi_J(t))'$, $\bar{\phi}^{(J)}(t) = \phi^{(J)}(t) - \mathbb{E} \{ \phi^{(J)}(v_i) \}$, $\phi_t^{(J)}(t) = (\phi_{1t}(t), \dots, \phi_{Jt}(t))'$. We approximate $\varsigma(t)$ by least squares projection on $\bar{\phi}^{(J)}(t)$. Denote the coefficients in this population projection by $a^{(J)}$. Then integration-by-parts leads to their identification by $a^{(J)} = [\mathbb{E} \{ \bar{\phi}^{(J)}(v_i) \bar{\phi}^{(J)}(v_i)' \}]^{-1} \mathbb{E} \{ \phi_{v_i}^{(J)}(v_i) \}$. On other hand, we can write the sample equivalent

$$\varsigma(t_i) = \Phi^{(J)}(t_i)' a^{(J)} + u_i, \quad i = 1, \dots, n, \quad (5.1)$$

for some zero mean, uncorrelated and homoskedastic (for ease of exposition we take their variance to be unity again) random variables u_i that are independent of elements of y_n , and $\Phi^{(J)}(t_i) = \phi^{(J)}(t_i) - n^{-1} \sum_{j=1}^n \phi^{(J)}(t_j)$.

For an observable vector $e = (e_1, \dots, e_n)'$, approximate $a^{(J)}$ by $a^{\dagger(J)}(e)$, where for generic $t = (t_1, \dots, t_n)'$, $a^{\dagger(J)}(t) = W^{(J)}(t)^{-1} w^{(J)}(t)$ with $W^{(J)}(t) = n^{-1} \sum_{i=1}^n \Phi^{(J)}(t_i) \Phi^{(J)}(t_i)'$, $w^{(J)}(t) = n^{-1} \sum_{i=1}^n \phi_{t_i}^{(J)}(t_i)$. The adaptive estimation literature then approximates $\varsigma(v_i)$ by $\Phi^{(J)}(e_i)' a^{\dagger(J)}$ and inserts this in the Newton step for estimating β . Our interest in this paper lies in testing the specification of $\varsigma(\cdot)$, not the Newton step in which an estimate based on this specification is inserted. If there are indeed increasingly many nonzero elements in $a^{(J)}$ in (5.1), the adaptive estimation methodology will be efficient. On the other hand if there are only a finite number of nonzero elements in $a^{(J)}$, $\varsigma(t)$ will almost have a parametric form. By simply testing the significance of coefficients in the first stage of adaptive estimation for a range of J , the practitioner can employ a better specification and anticipate better performance of estimates. Thus we treat (5.1) like a linear regression

model with increasingly many parameters and conduct tests of significance on increasingly large subvectors of $a^{(J)}$ using \mathbf{LM}_n , \mathbf{W}_n or \mathbf{LR}_n . For example, if we take $\phi_\ell(t) = t^\ell$ and cannot reject the null hypothesis that all except the leading coefficient in (5.1) are zero then $\varsigma(t)$ is just the score function of a standard normal distribution. As in Section 4 our tests can also be interpreted as providing a way to choose J in practice. Writing $u = (u_1, \dots, u_n)'$, $\bar{W}^{(J)}(t)$ for the $n \times J$ matrix with typical row $\Phi^{(J)}(t_i)$ and taking $\psi(x) = x^2/2$, we have $g_n = -n^{-1}\bar{W}'^{(J)}u$ (suppressing reference to the argument) and $H_n = n^{-1}\bar{W}'^{(J)}\bar{W}^{(J)} = W^{(J)}$, so here $M_n = \bar{W}^{(J)}$. Assumption A* in the proposition below defines the restrictions on J . It is quite technical and a repetition of conditions in Robinson (2005, 2010), so we state it in Appendix A. Simpler but stronger conditions restricting J are given in Newey (1988).

Proposition 5.1. *Suppose that we are testing the significance of a $\mathfrak{t}(J) \times 1$ subvector of $a^{(J)}$, where $\mathfrak{t}(J) \rightarrow \infty$ as $n \rightarrow \infty$, u_i satisfy the properties in Assumption 2, X_n is formed of uniformly bounded constants and Assumption 3 holds, Assumption A* holds and J is a function of n satisfying (3.7) with m constant, $s = J$, $q = \mathfrak{t}(J)$. Then \mathbf{LM}_n , \mathbf{W}_n and \mathbf{LR}_n have Property \mathcal{C} .*

Testing nonparametric density specifications against parametric alternatives through simple inference on series coefficients is more widely applicable. Following Gallant and Nychka (1987), there is a very large literature on maximum likelihood estimation via series approximations to smooth unknown densities. A detailed treatment is beyond the scope of this paper, but to name one example Gurmu, Rilstone, and Stern (1999) consider semiparametric estimation of a count regression model based on a series expansion of the unknown density of unobserved heterogeneity. The approach of this paper is likely to be extendable to obtain density specification tests with Property \mathcal{C} in this setting.

Consequences of misspecification in finite samples have effectively already been examined in a Monte Carlo study in Robinson (2010). In his design (cf. pg. 12 of that paper), adaptive estimates perform worse than the initial OLS for $J \geq 1$ when the true density of v is standard normal, as expected. However there are no remedies in that paper to correct or test for the specification as we have proposed.

6 Monte Carlo

Finite sample implications of the theory were examined in a set of Monte Carlo experiments. The first of these analysed the performance of \mathbf{LM}_n , \mathbf{W}_n and \mathbf{LR}_n when y_n was generated

using (2.6), (2.9) and (2.11). The aim is to assess quality of inferences for small to moderate sample sizes and fairly large parameter spaces, and not to choose sample sizes so large and parameter spaces so small that the experiments become uninformative. The following designs were used to generate $y \equiv y_n$ in each of the 5000 replications:

$$(2.6) : k = 20, 30, 40, q = 10, 15, 18, n = 150, 350, 700, \beta_0 = \iota_k/2, X \sim U(0, 1) \text{ but with first column of ones, } v \sim \mathcal{N}(0, 1),$$

$$(2.9) : k_1 = k_2 = 2, q = T = 5, 10, 15, N = 50, 100, 200, (\alpha_0, \beta_{10}, \beta_{20}) = (\iota_T, \iota_2/2, \iota_{2T}/2), X \sim U(0, 5), v \sim \mathcal{N}(0, 1),$$

$$(2.11) : k = 2, q = p = 8, 16, 32, m = 12, 48, 96, (\lambda_0, \beta_0) = (U(0, 1)_p, 1, 1/2), X \sim U(0, 5),$$

$$W_{in} = \text{diag} \left[0, \dots, \underbrace{(m-1)^{-1} (\iota_m \iota'_m - I_m)}_{i\text{-th diagonal block}}, \dots, 0 \right], v \sim \mathcal{N}(0, 1).$$

The unusual notation in the choice of λ_0 indicates that the λ_{0i} were generated from $U(0, 1)$ once at the start of the experiment (and then kept fixed for all combinations of p and m) to conform to a sufficient condition for the existence of a power series for S_n^{-1} viz. $|\lambda_{0i}| < 1, i = 1, \dots, p$ (cf. Proposition 2.1 of Gupta and Robinson (2015a)). They are: 0.81, 0.91, 0.13, 0.91, 0.63, 0.1, 0.28, 0.55, 0.96, 0.96, 0.16, 0.97, 0.96, 0.49, 0.8, 0.14, 0.42, 0.92 0.79, 0.96, 0.66, 0.04, 0.85, 0.93, 0.68, 0.76, 0.74, 0.39, 0.66, 0.17, 0.71, 0.03.

Because $\mathbb{E}(W_{in}y) = \mathbb{E}(W_{in}S_n^{-1}X\beta_0)$ this allows instruments to be taken as linearly independent columns of $W_{in}^r X, r \geq 1$. We maintain $r = 1$ in our experiments, as is common in the SAR literature, for a total of kp instruments apart from those in X , and analyse only IV estimates. The choice of W_{in} is commonly employed in Monte Carlo simulations using the SAR model and comes from Case (1991, 1992), who models an economy in which there are p districts each with m farmers who influence each farmer in their own district equally and are independent of farmers in other districts.

We first report empirical sizes and powers in Tables 6.1 and 6.2 for the following nulls:

$$(2.6) : H_0 : \beta = \beta_0 = \iota_k, H_0 : \beta = 0.4\iota_k,$$

$$(2.9) : H_0 : \alpha = \alpha_0 = \iota_T, H_0 : \alpha = 0.45\iota_T,$$

$$(2.11) : H_0 : \lambda = \lambda_0 = U(0, 1)_p, H_0 : \lambda_i = \lambda_j, i \neq j, i, j = 1, \dots, p.$$

Tests of the type we construct are one-sided because only positive increasing values occur asymptotically under the alternative hypothesis, so we employ one-sided critical values to

(2.6)	n	150			350			700		
k	q	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
20	10	0.0820	0.0912	0.0834	0.0732	0.0756	0.0806	0.0650	0.0664	0.0724
	15	0.0886	0.0752	0.0908	0.0734	0.0684	0.0874	0.0672	0.0652	0.0856
	18	0.0872	0.0596	0.0952	0.0688	0.0574	0.0864	0.0674	0.0630	0.0916
30	10	0.0736	0.1150	0.0752	0.0800	0.0932	0.0852	0.0684	0.0760	0.0800
	15	0.0794	0.1014	0.0792	0.0766	0.0848	0.0880	0.0690	0.0714	0.0866
	18	0.0796	0.0900	0.0840	0.0762	0.0788	0.0886	0.0690	0.0704	0.0946
40	10	0.0836	0.1684	0.0792	0.0698	0.1026	0.0724	0.0660	0.0768	0.0756
	15	0.0894	0.1572	0.0856	0.0664	0.0892	0.0790	0.0662	0.0762	0.0862
	18	0.0914	0.1530	0.0892	0.0654	0.0876	0.0818	0.0672	0.0768	0.0902
(2.9)	N	50			100			200		
	T	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
	5	0.1096	0.0810	0.0680	0.0900	0.0776	0.0704	0.0738	0.0654	0.0636
	10	0.1256	0.0838	0.0766	0.0894	0.0714	0.0732	0.0752	0.0682	0.0762
	15	0.1374	0.0858	0.0850	0.0928	0.0696	0.0886	0.0886	0.0776	0.0922
(2.11)	m	12			48			96		
	p	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
IV	8	0.0914	0.2746	0.2036	0.0694	0.1058	0.0992	0.0704	0.0830	0.0856
	16	0.0900	0.2928	0.2226	0.0730	0.1058	0.1120	0.0660	0.0794	0.0958
	32	0.0884	0.4180	0.2838	0.0692	0.1230	0.1212	0.0650	0.0842	0.1006

Table 6.1: Monte Carlo size for OLS for (2.6) and (2.9), IV for (2.11). Nominal size is 5%.

compute the values in the tables. This was noted also by Hong and White (1995).

For (2.6) all three tests tend to be oversized compared to the nominal 5%, but converging towards the latter as n increases even though the improvement is not always monotonic, as seen by the behaviour of \mathbf{LR}_n for $q = 18$. Over-sizing worsens for all three tests as k increases for given n and also worsens for \mathbf{W}_n and \mathbf{LR}_n as q increases for given n and k but improves for \mathbf{LM}_n , again with occasional exceptions to monotonicity. This reflects the fact that \mathbf{LM}_n relies on estimation of the model under the null, which is more parsimonious, while the other two tests also require estimation of the unrestricted model. The powers for (2.6) displayed in the top panel of Table 6.2 indicate improvement again for larger samples in all tests. However power increases with q (given n, k) for \mathbf{W}_n and \mathbf{LR}_n but decreases for \mathbf{LM}_n , with similar justification of this behaviour as provided for the sizes. Increase in k (given n) doesn't seem to alter the behaviour of any test in a discernable pattern.

With (2.9) note that n increases with T , which is also the rate of increase of the

(2.6)	n	150			350			700		
k	q	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
20	10	0.1290	0.1404	0.1160	0.1952	0.2008	0.1890	0.3596	0.3628	0.3580
	15	0.1396	0.1200	0.1270	0.2304	0.2208	0.2268	0.4544	0.4456	0.4536
	18	0.1488	0.1088	0.1316	0.2564	0.2298	0.2470	0.4920	0.4814	0.4926
30	10	0.1172	0.1684	0.0990	0.1962	0.2268	0.1936	0.3516	0.3682	0.3542
	15	0.1340	0.1686	0.1138	0.2270	0.2432	0.2164	0.4338	0.4436	0.4338
	18	0.1426	0.1560	0.1204	0.2456	0.2540	0.2382	0.4792	0.4822	0.4786
40	10	0.1204	0.2298	0.1038	0.1954	0.2448	0.1928	0.3490	0.3812	0.3510
	15	0.1324	0.2334	0.1092	0.2258	0.2716	0.2176	0.4408	0.4682	0.4372
	18	0.1500	0.2288	0.1192	0.2426	0.2820	0.2290	0.4804	0.5048	0.4822
(2.9)	N	50			100			200		
	T	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
	5	0.5478	0.4836	0.4568	0.8076	0.7798	0.7746	0.9854	0.9838	0.9852
	10	0.7490	0.6780	0.6520	0.9560	0.9464	0.9426	0.9996	0.9996	0.9998
	15	0.8620	0.7964	0.7696	0.9912	0.9888	0.9872	1.0000	1.0000	1.0000
(2.11)	m	12			48			96		
	p	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n	\mathbf{W}_n	\mathbf{LM}_n	\mathbf{LR}_n
IV	8	0.1470	0.8198	1.0000	0.9314	0.9992	1.0000	1.0000	1.0000	1.0000
	16	0.1456	0.9654	1.0000	0.9966	1.0000	1.0000	1.0000	1.0000	1.0000
	32	0.1348	0.9986	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6.2: Monte Carlo power for OLS for (2.6) and (2.9), IV for (2.11).

parameter space. There is over-sizing with all three tests, but \mathbf{LR}_n is the closest to the nominal 5%. The over-sizing worsens with increasing T for given N but not in each case, e.g. \mathbf{LM}_n when $N = 100$ actually improves with increasing T . The conclusion is that large T can have serious consequences in inference on T fixed-effect parameters even though it implies a larger $n = NT$. On the other hand, powers displayed in Table 6.2 for all three tests improve in all possible ways that n can increase, viz. both N and T increase (diagonal), only N increases (horizontal), only T increases (vertical). Unit power is attained by all three for the largest sample with $T = 15, N = 200$.

Finally, for (2.11) recall that n increases with p , which is also the rate of increase of the parameter space. We find that \mathbf{LM}_n and \mathbf{LR}_n are unacceptably oversized for $m = 12$, with \mathbf{W}_n much better. However matters do improve for \mathbf{LM}_n and \mathbf{LR}_n as m increases, though the improvement is not necessarily better with both m and p increasing (so both parameter space and sample size increase) than with m increasing for given p . In fact

sizes are usually worse in the former case indicating that the gains for \mathbf{LM}_n and \mathbf{LR}_n due to increased sample size are overpowered in this case by the burden of estimating extra parameters. On the other hand \mathbf{W}_n showcases better performance and improves more as we proceed diagonally on the table as opposed to horizontally, reflecting a better response to increasing n with both m and p . The powers in Table 6.2 tell a somewhat different story, with \mathbf{W}_n under-performing \mathbf{LM}_n and \mathbf{LR}_n substantially when $m = 12$ but all tests giving excellent results for larger m . \mathbf{LR}_n seems to be the clear winner here, always giving unit power, while the latter property is true for all three tests when $m = 96$.

Our second set of experiments pertain to the specification testing procedure described in Section 4, using the polynomial basis described there. In each of 5000 replications we generate y_i , $i = 1, \dots, n$ and $n = 100, 300, 500$, using several data generating processes (DGPs):

$$\begin{aligned} DGP1 : y_i &= 1 + x_{1i} + x_{2i} + v_i \equiv x_i' \beta_0 + v_i, \\ DGP2(\tau_1) : y_i &= \exp \{ \tau_1 (x_i' \beta_0) \} + v_i, \\ DGP3(\tau_2) : y_i &= \tau_2 \log \{ \log (x_i' \beta_0) \} + v_i, \\ DGP4(\tau_3) : y_i &= 1 + \tau_3 (\sin x_{1i} + \cos x_{2i}) + v_i, \\ DGP5(\tau_4) : y_i &= 1 + \tau_4 \left(x_{1i}^{-2} + x_{2i}^{-\frac{1}{2}} \right) + v_i, \end{aligned}$$

where $x_{1i} = (z_i + z_{1i})/2$, $x_{2i} = (z_i + v_{zi})/2$ with $z_i, z_{1i}, z_{2i} \sim \text{iid } U(0, 5)$, $v_i \sim \text{iid } \mathcal{N}(0, 1)$ and τ_i , $i = 1, 2, 3, 4$, some real numbers. Note that we have redefined $\beta_0 = \iota_3$.

We experimented with $J = 1, 2, 4$. (4.12) and the discussion in the paragraph below suggests an upper bound choice for J of $\left\lceil n^{\frac{1}{6}} \right\rceil$, where we take $\lceil x \rceil$ to denote the integer part of x . This is the kind of approach used by Hong and White (1995) when choosing J in their simulations (they use closest integer, not integer part), but Robinson (2005, 2010) criticizes reliance on an upper bound for choice of J , stating that asymptotic theory provides little guidance. Our discussion of simulation results for $DGP5$ below give credence to this criticism. Our choices of n imply $\left\lceil n^{\frac{1}{6}} \right\rceil = [2.15], [2.59], [2.82] = 2, 2, 2$, but we choose to also report for $J = 4$ to give a sharp illustration of the consequences of ‘overfitting’ the upper bound. For $DGP5$ we find that power improves, thus indicating that the upper bound may not always be the optimal choice.

When $J = 1$ we are simply regressing y on a constant, x_1 and x_2 , and under the null we set both slope coefficients equal to zero. Thus the tests boil down to a test of significance of x_1 and x_2 , and rejection probabilities are interpreted as power against an alternative that

one of x_1 or x_2 is significant. For $J = 2, 4$ the null hypothesis sets all coefficients apart from those on $(1, x_1, x_2)$ to be zero. This null is true under *DGP1* so the rejection probabilities under this DGP are to be interpreted as sizes to be compared to the nominal 5%, while under the other four DGPs the null is not true and the rejection probabilities are to be interpreted as power of a null hypothesis of linearity against these DGPs as alternatives.

The rejection probabilities are tabulated in Table 6.3. Under *DGP1*, the first row indicates unit power for all three tests and sample sizes. For $J \geq 2$, the next two rows display empirical sizes. All three tests are over-sized, but acceptable for $n = 500$ while even for $n = 300$ they are not very far off with the best being \mathbf{LR}_n with $J = 2$. There is no clear winner between $J = 2$ and $J = 4$. Next we take $\tau_1 = 0.3, 0.2$ in *DGP2*(τ_1). Note that the expansion $\exp(x) = \sum_{j=0}^{\infty} x^j/j!$ indicates that the smaller the absolute value of τ_1 the closer *DGP*(τ_1) is to *DGP1*. Thus tests against linearity are expected to have more power for larger τ_1 , which in fact should also lead to better power in tests of regressor significance. As discussed above our tests boil down to the latter when $J = 1$, and in this case we see that when $\tau_1 = 0.3$ we get unit power always while for $\tau_1 = 0.2$ the power is still excellent but not quite unity for $n = 100$. The power under a null of linearity is always better for $J = 2$ than for $J = 4$, although for both choices it increases with n . With $J = 2$ and $\tau_1 = 0.3$, it starts at below 30% for all three tests when $n = 100$ but improves to over 70% when $n = 300$ and is around 90% for $n = 500$. On the other hand the tests lose power quite dramatically for $\tau_1 = 0.2$, doing no better than 15.4%, although they still improve with increasing n . Next we take $\tau_2 = 1, 0.8$ for *DGP3*(τ_2), finding rejection percentages when $J = 1$ to be excellent for both cases with unit power achieved when $n = 500$, nearly unit power for $n = 300$. For $n = 100$ power is much better for $\tau_2 = 1$. Power under the null of linearity follows much the same pattern, and like *DGP2*(τ_1) power for $J = 2$ dominates that for $J = 4$. With both choices of τ_2 power triples from $n = 100$ to $n = 500$, but from around 25% to around 75.5% when $\tau_2 = 1$ and from around 19% to around 57% when $\tau_2 = 0.8$. All three tests have very similar performance.

Next we study *DGP4*(τ_3) with $\tau_3 = 0.5, 0.2$. Power when the null imposes insignificance of the regressors becomes unity for $\tau_3 = 0.5$, $n \geq 300$, although it is over 95% even when $n = 100$. With $\tau_3 = 0.2$, power is lower for $n = 100$ but still very good for $n = 500$. When the null is of linearity, $J = 4$ has better power properties than $J = 2$ for both values of τ_3 but the power is much higher for the bigger value. An almost analogous analysis holds for *DGP5*(τ_4), which we simulate with $\tau_4 = 0.3, 0.1$. But, as mentioned earlier, there is one crucial difference: this is the only nonlinear DGP for which $J = 4$ dominates $J = 2$, and

in fact it does so for all tests, sample sizes and both values of τ_4 .

7 Empirical illustrations

Data for both examples is available at <https://www.economics.utoronto.ca/yatchew/>.

7.1 Scale economies in electricity distribution

Yatchew (2000) considers (4.10) as a variant of the Cobb-Douglas model for the costs of distributing electricity with data on 81 municipal distributors in Ontario, Canada, during 1993. The interest lies in examining economies of scale in the number of customers. He takes $y = tc$, $x_1 = (wage, pcap, PUC, kwh, life, lf, kmwire)'$ and $x_2 = cust$, where tc is the log of total cost per customer, $wage$ is the log wage rate, $pcap$ is the log price of capital, PUC is a dummy variable for public utility commissions that deliver additional services, kwh is the log of kilowatt hour sales per customer, $life$ is the log of remaining life of distribution assets, lf is the log of the load factor (which measures capacity utilization relative to peak usage), $kmwire$ is the log of kilometres of distribution wire per customer and $cust$ is the log of the number of customers. Yatchew (2003) also fits a fully parametric specification with $d(cust) = \beta_1 cust + \beta_2 cust^2$. He uses a differencing procedure and his test (pg. 9) fails marginally to reject quadraticity, obtaining a test statistic of 1.5 to be compared to the 5% standard normal critical value of 1.645. However he later (pg. 77) employs a different specification test, also asymptotically standard normal, and finds that quadraticity is rejected with a statistic of 2.4. We will employ \mathbf{W}_n to test for a quadratic specification, by fitting (4.10) using the series $\sum_{j=1}^J \alpha_{0j} cust^j$ with $J = 4, 5$.

When $J = 4$, $q = 2$ and the null is $\alpha_{03} = \alpha_{04} = 0$ while for $J = 5$, $q = 3$ and the null is $\alpha_{03} = \alpha_{04} = \alpha_{05} = 0$. We get $\mathbf{W}_n = 0.1, 3.51$ for $J = 4, 5$ respectively. Compared to the one-sided 5% critical value of 1.645 for a standard normal distribution this results in a rejection of the null hypothesis of quadraticity when $J = 5$. Because of the very small statistic when $J = 4$ we conclude that including high order polynomial terms captures features of customer scale economies in electricity distribution that might otherwise be missed, thus providing evidence in favour of a semiparametric specification.

7.2 Engel curve estimation: testing for linearity and equality

Yatchew (2003) considers Engel curve estimation from South African household survey

data. Two categories are considered: single individuals without children (‘singles’, 1,109 observations) and couples without children (‘couples’, 890 observations). For each group we estimate (4.1) with $y = fs$ and $x = \text{texp}$ where fs is the food share of total expenditure and texp is the log of total expenditure, using the series $\sum_{j=1}^J \alpha_{0j} \text{texp}^j$ with $J = 2, 3, 4$, plotting the results in Figure 7.1. One question of interest centres on whether the Engel curves are linear, to answer which we test $\alpha_{02} = 0$, $\alpha_{02} = \alpha_{03} = 0$, $\alpha_{02} = \alpha_{03} = \alpha_{04} = 0$ for $J = 2, 3, 4$, respectively. Let s and c superscripts denote singles and couples respectively. We obtain $\mathbf{W}_n^s = 10.47$; $\mathbf{W}_n^c = 1.83$, for $J = 2$, $\mathbf{W}_n^s = 13.18$; $\mathbf{W}_n^c = 17.07$, for $J = 3$ and $\mathbf{W}_n^s = 14.13$; $\mathbf{W}_n^c = 13.98$, for $J = 4$, indicating a rejection of the linear specification on comparison with a critical value of 1.645. We thus opt for a nonparametric model with $J = 4$, because taking $J > 4$ leads to multicollinearity problems.

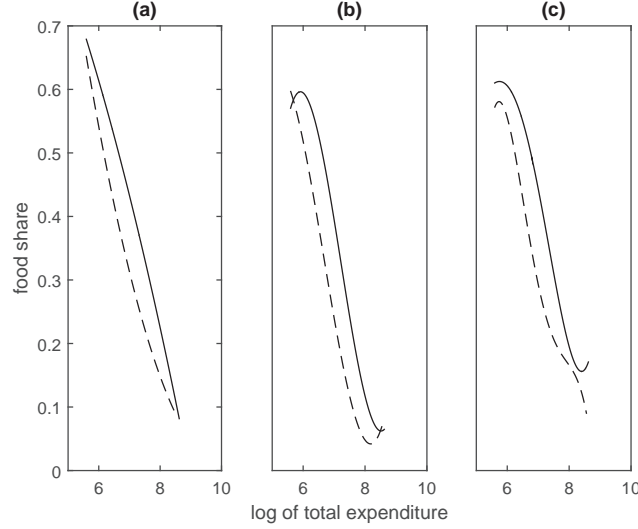


Figure 7.1: Fitted Engel curve for ‘singles’ (dashed line) and ‘couples’ (solid line), with $J = 2$ (a), $J = 3$ (b) and $J = 4$ (c).

Once we have chosen our model (with $J = 4$), an economic question of interest is: if the two Engel curves in Figure 7.1(c) were to be superimposed, would they coincide? It looks unlikely given the marked difference in shapes for high levels of texp . We answer this question by generating a dummy variable *coup* that takes the value 1 for couples and 0 otherwise, pooling the data for singles and couples, estimating the model

$$fs = \alpha_{00} + \sum_{j=1}^4 \alpha_{0j} \text{texp}^j + \sum_{j=1}^4 \delta_{0j} (\text{coup} \times \text{texp}^j) + v_i, \quad (7.1)$$

and testing $\delta_{01} = \delta_{02} = \delta_{03} = \delta_{04} = 0$. This returns $\mathbf{W}_n = 88.87$, implying a strong rejection of the null hypothesis. In contrast the test used in Yatchew (2003) (cf. pg. 69) returned a statistic of 1.76 and therefore a rather weak rejection of the null of equality.

8 Conclusion and robustification to heteroskedasticity

We have proposed nonparametric specification tests based on the trinity of testing principles for models in which objective function derivatives may be linear in iid disturbances. The test statistic sequences were shown to be asymptotically standard normal under the null hypothesis, consistent and possessing nontrivial power against local alternatives converging to the null at a prescribed rate. We now briefly discuss heteroskedasticity robust versions of the tests, in particular considering \mathbf{LM}_n and \mathbf{W}_n statistics. Suppose that the unit variance assumption on the ϵ_i is removed and instead $\text{var}(\epsilon_i) = \sigma_i^2$, $i = 1, \dots, n$, where $\sigma_i^{-2} \leq C$. Note also that the existence of finite fourth moments guarantees that $\sigma_i^2 \leq C$. Because so many of our cases of interest essentially involve only linear regression, we discuss heteroskedasticity robust versions of \mathbf{LM}_n and \mathbf{W}_n in (2.6). As for the homoskedastic case, these are obtained simply by standardizing usual heteroskedasticity robust versions of familiar statistics. Define $\hat{\epsilon} = (\hat{\epsilon}_1, \dots, \hat{\epsilon}_n)'$, $\tilde{\epsilon} = (\tilde{\epsilon}_1, \dots, \tilde{\epsilon}_n)'$, $\hat{\epsilon}_i = y_i - x_i' \hat{\beta}$, $\tilde{\epsilon}_i = y_i - x_i' \tilde{\beta}$, $\hat{\Omega} = \text{diag}[\hat{\epsilon}_1^2, \dots, \hat{\epsilon}_n^2]$, $\tilde{\Omega} = \text{diag}[\tilde{\epsilon}_1^2, \dots, \tilde{\epsilon}_n^2]$ and

$$\begin{aligned} \mathbf{LM}_n^{\mathcal{R}} &= \frac{\hat{\epsilon}' \hat{\Omega}^{-1} X_1 \left\{ X_1' \hat{\Omega}^{-1} X_1 - X_1' \hat{\Omega}^{-1} X_2 \left(X_2' \hat{\Omega}^{-1} X_2 \right)^{-1} X_2' \hat{\Omega}^{-1} X_1 \right\}^{-1} X_1' \hat{\Omega}^{-1} \hat{\epsilon} - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}}, \\ \mathbf{W}_n^{\mathcal{R}} &= \frac{\tilde{\beta}_1' \left\{ X_1' \tilde{\Omega}^{-1} X_1 - X_1' \tilde{\Omega}^{-1} X_2 \left(X_2' \tilde{\Omega}^{-1} X_2 \right)^{-1} X_2' \tilde{\Omega}^{-1} X_1 \right\} \tilde{\beta}_1 - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}}, \end{aligned}$$

where $X = [X_1, X_2]$ is a partition conformable to the dimension of the null.

n		100			300			500		
	J	W_n	LM_n	LR_n	W_n	LM_n	LR_n	W_n	LM_n	LR_n
$DGP1$	1	100	100	100	100	100	100	100	100	100
	2	8.6	8.5	8.1	7.6	7.6	7	6.3	6.3	6.2
	4	8.9	7	7.4	8	7.3	8.8	6.4	5.8	7.9
$DGP2(0.3)$	1	100	100	100	100	100	100	100	100	100
	2	28.6	28.2	27.5	71.2	71	71.1	89.9	89.9	90.2
	4	18.9	15.1	16	47.8	45.6	47.2	67.9	67.3	68.7
$DGP2(0.2)$	1	96.3	96.1	96.5	100	100	100	100	100	100
	2	9.6	9.4	8.8	12.6	12.4	12.5	15.4	15.2	15.3
	4	9.7	7.6	8.4	9.8	8.9	10	10.1	9.5	10.9
$DGP3(1)$	1	80.1	79.1	81.4	99.9	99.9	99.9	100	100	100
	2	25.5	24.9	24.7	51	50.9	51.7	75.6	75.4	76.1
	4	20.2	17	18.1	40.5	38.5	41.7	65.7	64.9	65.8
$DGP3(0.8)$	1	63.7	62.9	63.9	97.5	97.5	97.4	100	100	100
	2	19	18.2	17.7	36.7	36.4	36.6	57.2	57.1	57.3
	4	15.9	12.2	14.1	27.4	25.6	26.8	44.9	44.1	44.2
$DGP4(0.5)$	1	95.5	95.1	96.2	100	100	100	100	100	100
	2	43.1	42.5	42.8	88.3	88.2	88.8	98.6	98.6	98.7
	4	31.7	26.6	30.1	77.6	76.4	78.7	95.6	95.3	95.9
$DGP4(0.2)$	1	34	32.7	33.4	75	74.8	75.3	92.6	92.6	92.9
	2	13.5	13.1	12.6	24.4	24.3	24.5	36.4	36.3	35.9
	4	11.6	8.7	10.8	17	15.8	17.7	25.4	24.5	25.4
$DGP5(0.3)$	1	60.4	59.4	64.5	95.9	95.9	96.5	99.7	99.7	99.8
	2	55.3	55	57.6	91.7	91.6	92.4	98.6	98.5	98.7
	4	60.2	57.6	60.5	94.5	94.1	94.8	99.6	99.6	99.6
$DGP5(0.1)$	1	22.1	21.5	25.6	46.7	46.5	51.4	62.2	62	66
	2	25	24.8	26.7	49.8	49.6	51.9	65.9	65.8	67.3
	4	30.4	27.4	29.8	60.5	59.5	61	77	76.4	77.4

Table 6.3: Comparison of specification tests: Rejection percentages

Appendices

A Proofs of theorems

Proof of Theorem 3.1. We first carry out some preliminary development. Because $\theta_{10} = 0$ under H_0 , by the mean value (MVT) theorem we can write $\hat{g}_{1,n} = g_{1,n} + \bar{H}_{12,n} \Delta_{\theta_{20}}^{\hat{\theta}_2}$, $0 = \hat{g}_{2,n} = g_{2,n} + \bar{H}_{22,n} \Delta_{\theta_{20}}^{\hat{\theta}_2}$ with $\|\Delta_{\theta_0}^{\bar{\theta}}\| \leq \|\Delta_{\theta_0}^{\hat{\theta}}\|$, whence

$$\hat{g}_{1,n} = \mathcal{X}_n^{\bar{H}} [g'_{1,n}, g'_{2,n}]' = \mathcal{X}_n^{\bar{H}} g_n. \quad (\text{A.1})$$

Note that the value of $\bar{\theta}$ may be different in each row of $\bar{H}_{12,n}$, this fact applying whenever the MVT is applied to a vector of values. Then (3.1) becomes

$$\frac{ng'_n \mathcal{H}_n g_n - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} \quad (\text{A.2})$$

where $\mathcal{H}_n = \mathcal{X}_n^{\bar{H}} \hat{H}_n^{11} \mathcal{X}_n^{\bar{H}}$. Assumption 2 and Lemma B.5 imply that (A.2) is $\sum_{i=1}^3 \Pi_{i,n}$ with

$$\Pi_{1,n} = \frac{n \Delta_{n^{-1}M'\epsilon}^g \mathcal{H}_n \Delta_{n^{-1}M'\epsilon}^g}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = \mathcal{O}_p \left(n \|\Delta_{n^{-1}M'\epsilon}^g\|^2 / q^{\frac{1}{2}} \right), \quad (\text{A.3})$$

$$\Pi_{2,n} = 2 \frac{\epsilon' M_n \mathcal{H}_n \Delta_{n^{-1}M'\epsilon}^g}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = \mathcal{O}_p \left(\|\Delta_{n^{-1}M'\epsilon}^g\| \|M'_n \epsilon\| / q^{\frac{1}{2}} \right), \quad (\text{A.4})$$

$$\Pi_{3,n} = \frac{n^{-1} \epsilon' M_n \mathcal{H}_n M'_n \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}}. \quad (\text{A.5})$$

(A.3), (A.4) are negligible by (3.4) so (A.2) is $\Pi_{3,n} + o_p(1)$. Denoting $\xi_n = \mathcal{X}_n^{\bar{H}} M'_n \epsilon$ we have $\Pi_{3,n} = (n^{-1} \xi'_n \hat{H}_n^{11} \xi_n - q) / 2^{\frac{1}{2}} q^{\frac{1}{2}}$. Note that by the proof of Lemma B.5, we have $\xi_n = \mathcal{O}_p(\|M'_n \epsilon\|)$. We write $\Pi_{3,n} = \sum_{i=1}^4 \Gamma_{3i,n}$ with

$$\Gamma_{31,n} = \frac{n^{-1} \xi'_n \Delta_{H^{11}}^{\hat{H}} \xi_n}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = \mathcal{O}_p \left(n^{-1} \|M'_n \epsilon\|^2 \|\Delta_{H^{\hat{H}}}\| / q^{\frac{1}{2}} \right), \quad (\text{A.6})$$

$$\Gamma_{32,n} = \frac{n^{-1} \xi'_n \Delta_{n^{-1}M'M^{11}}^{H^{\hat{H}}} \xi_n}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = \mathcal{O}_p \left(n^{-1} \|M'_n \epsilon\|^2 \|\Delta_{n^{-1}M'M}^H\| / q^{\frac{1}{2}} \right), \quad (\text{A.7})$$

$$\Gamma_{34,n} = \frac{n^{-1} \xi'_n (n^{-1} M'_n M_n)^{11} \xi_n - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}}. \quad (\text{A.8})$$

The bound for the $\Delta_{H^{11}}^{\hat{H}^{11}}$ term in (A.6) is justified by the following argument. First $\left\| \Delta_{H^{11}}^{\hat{H}^{11}} \right\| \leq \|H_n^{11}\| \left\| \hat{H}_n^{11} \right\| \left\| \Delta_{(H^{11})^{-1}}^{(\hat{H}^{11})^{-1}} \right\|$, with the first and second factors on the RHS $\mathcal{O}_p(1)$ by Lemma B.2. By the partitioned inverse formula $\left(\hat{H}_n^{11} \right)^{-1} = \hat{H}_{11,n} - \hat{H}_{12,n} \hat{H}_{22,n}^{-1} \hat{H}_{21,n}$ and $(H_n^{11})^{-1} = H_{11,n} - H_{12,n} H_{22,n}^{-1} H_{21,n}$, implying that the last factor is bounded by $\left\| \Delta_{H_{11,n}}^{\hat{H}_{11,n}} \right\| + \left\| \Delta_{H_{12,n} H_{22,n}^{-1} H_{21,n}}^{\hat{H}_{12,n} \hat{H}_{22,n}^{-1} \hat{H}_{21,n}} \right\|$. The first term in the last expression is bounded by $\left\| \Delta_H^{\hat{H}} \right\|$, while the second term is bounded by $\left\| \hat{H}_{12,n} \right\| \left\| \hat{H}_{21,n} \right\| \left\| H_{22,n}^{-1} \right\| \left\| \hat{H}_{22,n}^{-1} \right\| \left\| \Delta_{H_{22,n}}^{\hat{H}_{22,n}} \right\| + \left\| H_{22,n}^{-1} \right\| \left\| \hat{H}_{21,n} \right\| \left\| \Delta_{H_{12,n}}^{\hat{H}_{12,n}} \right\| + \left\| H_{22,n}^{-1} \right\| \left\| H_{12,n} \right\| \left\| \Delta_{H_{21,n}}^{\hat{H}_{21,n}} \right\|$. By Lemmas B.2 and B.3, we conclude that $\left\| \hat{H}_{12,n} \right\|$, $\left\| \hat{H}_{21,n} \right\|$, $\left\| H_{12,n} \right\|$, $\left\| \hat{H}_{22,n}^{-1} \right\|$ and $\left\| H_{22,n}^{-1} \right\|$ are all $\mathcal{O}_p(1)$. $\left\| \Delta_{H_{22,n}}^{\hat{H}_{22,n}} \right\|$, $\left\| \Delta_{H_{12,n}}^{\hat{H}_{12,n}} \right\|$ and $\left\| \Delta_{H_{21,n}}^{\hat{H}_{21,n}} \right\|$ are all bounded by $\left\| \Delta_H^{\hat{H}} \right\|$. The bound in (A.6) now follows, and the bound in (A.7) is justified similarly. By (3.4), (A.6), (A.7) are negligible. Finally we show that

$$\frac{n^{-1} \xi'_n (n^{-1} M'_n M_n)^{11} \xi_n}{2^{\frac{1}{2}} q^{\frac{1}{2}}} - \frac{n^{-1} \epsilon' \mathcal{M}_n \epsilon}{2^{\frac{1}{2}} q^{\frac{1}{2}}} = o_p(1). \quad (\text{A.9})$$

Adding and subtracting $n^{-1} \xi'_n (n^{-1} M'_n M_n)^{11} \mathcal{X}_n^{n^{-1} M' M} M'_n \epsilon / q^{\frac{1}{2}}$ to the LHS and combining terms yields

$$\frac{1}{2^{\frac{1}{2}} q^{\frac{1}{2}}} n^{-1} \epsilon' M_n \left(\mathcal{X}_n^{\bar{H}} + \mathcal{X}_n^{n^{-1} M' M} \right)' (n^{-1} M'_n M_n)^{11} \left[0, \Delta_{\bar{H}_{12} \bar{H}_{22}^{-1}}^{n^{-1} M' M_{12} n^{-1} M' M_{22}^{-1}} \right] M'_n \epsilon. \quad (\text{A.10})$$

$\left\| \Delta_{\bar{H}_{12} \bar{H}_{22}^{-1}}^{n^{-1} M' M_{12} n^{-1} M' M_{22}^{-1}} \right\| \leq \sum_{i=1}^4 \|\Upsilon_{i,n}\|$ with $\Upsilon_{1,n} = \Delta_{\bar{H}_{12}}^{\bar{H}_{12}} \bar{H}_{22,n}^{-1}$, $\Upsilon_{2,n} = \Delta_{n^{-1} M' M_{12}}^{H_{12}} \bar{H}_{22,n}^{-1}$, $\Upsilon_{3,n} = M' M_{12,n} \bar{H}_{22,n}^{-1} \Delta_{\bar{H}_{22}}^{H_{22}} H_{22,n}^{-1}$ and $\Upsilon_{4,n} = M' M_{12,n} (n^{-1} M'_n M_n)_{22}^{-1} \Delta_{H_{22}}^{n^{-1} M' M_{22}} H_{22,n}^{-1}$, so

$$\|\Upsilon_{1,n}\| \leq \left\| \Delta_{\bar{H}_{12}}^{\bar{H}_{12}} \right\| \left\| \bar{H}_{22,n}^{-1} \right\| = \mathcal{O}_p \left(\left\| \Delta_H^{\bar{H}} \right\| \right), \quad (\text{A.11})$$

$$\|\Upsilon_{2,n}\| \leq \left\| \Delta_{n^{-1} M' M_{12}}^{H_{12}} \right\| \left\| \bar{H}_{22,n}^{-1} \right\| = \mathcal{O}_p \left(\left\| \Delta_{n^{-1} M' M}^H \right\| \right), \quad (\text{A.12})$$

$$\|\Upsilon_{3,n}\| \leq \left\| M' M_{12,n} \right\| \left\| \bar{H}_{22,n}^{-1} \right\| \left\| H_{22,n}^{-1} \right\| \left\| \Delta_{\bar{H}_{22}}^{\bar{H}_{22}} \right\| = \mathcal{O}_p \left(\left\| \Delta_H^{\bar{H}} \right\| \right), \quad (\text{A.13})$$

$$\|\Upsilon_{4,n}\| \leq \left\| M' M_{12,n} \right\| \left\| H_{22,n}^{-1} \right\| \left\| (n^{-1} M'_n M_n)_{22}^{-1} \right\| \left\| \Delta_{n^{-1} M' M_{22}}^{H_{22}} \right\| = \mathcal{O}_p \left(\left\| \Delta_{n^{-1} M' M}^H \right\| \right), \quad (\text{A.14})$$

by Lemmas B.2, B.3, B.6 and proof of Lemma B.5. Thus, by Lemma B.6, (A.10) is

$$\mathcal{O}_p \left(n^{-1} \|M'_n \epsilon\|^2 \max \left\{ \left\| \Delta_H^{\bar{H}} \right\|, \left\| \Delta_{n^{-1} M' M}^H \right\| \right\} / q^{\frac{1}{2}} \right) = o_p(1),$$

proving (A.9). The theorem is now proved. \square

Proof of Theorem 3.2. By the MVT, $0 = \tilde{g}_n = g_n + \bar{H}_n \Delta_{\theta_0}^{\tilde{\theta}}$, with $\|\Delta_{\theta_0}^{\tilde{\theta}}\| \leq \|\Delta_{\theta_0}^{\tilde{\theta}}\|$, so solving yields $\tilde{\theta}_1 = -[\bar{H}_n^{11}, \bar{H}_n^{12}] g_n$. Thus, upon substituting in (3.2), \mathbf{W}_n is

$$\frac{n^{-1} \epsilon' M_n [n^{-1} M'_n M_n^{11}, n^{-1} M'_n M_n^{12}]' (n^{-1} M'_n M_n^{11})^{-1} [n^{-1} M'_n M_n^{11}, n^{-1} M'_n M_n^{12}] M'_n \epsilon - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} + \mathcal{O}_p \left(\left[\|\Delta_{n^{-1} M'_n \epsilon}^g\| \{n \|\Delta_{n^{-1} M'_n \epsilon}^g\| + \|M'_n \epsilon\|\} + n^{-1} \|M'_n \epsilon\|^2 \left\{ \|\Delta_H^{\tilde{H}}\| + \|\Delta_{n^{-1} M'_n M}^H\| \right\} \right] / q^{\frac{1}{2}} \right).$$

By (3.4) the second term above is negligible. The quadratic form in $M'_n \epsilon$ in the first term is weighted by a matrix of rank q , which by the partitioned matrix inversion formula simplifies to $(n^{-1} M'_n M_n)^{-1} - \Sigma_n^{n^{-1} M'_n M}$, whence the claim follows. Now consider \mathbf{LR}_n . By the MVT we have

$$\mathbf{LR}_n = \frac{n \Delta_{\theta}^{\hat{\theta}} \bar{H}_n \Delta_{\theta}^{\hat{\theta}} - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}}, \quad (\text{A.15})$$

$$0 = \tilde{g}_{1n} = g_{1n} + \bar{H}_{11,n} \tilde{\theta}_1 + \bar{H}_{12,n} \Delta_{\theta_{20}}^{\tilde{\theta}_2}, \quad (\text{A.16})$$

$$0 = \tilde{g}_{2n} = g_{2n} + \bar{H}_{21,n} \tilde{\theta}_1 + \bar{H}_{22,n} \Delta_{\theta_{20}}^{\tilde{\theta}_2}, \quad (\text{A.17})$$

$$0 = \hat{g}_{2n} = g_{2n} + \check{H}_{22,n} \Delta_{\theta_{20}}^{\hat{\theta}_2}, \quad (\text{A.18})$$

with $\|\Delta_{\theta_0}^{\tilde{\theta}}\| \leq \|\Delta_{\theta_0}^{\tilde{\theta}}\|$ and $\|\Delta_{\theta_0}^{\check{\theta}}\| \leq \|\Delta_{\theta_0}^{\hat{\theta}}\|$. Subtracting (A.18) from (A.17) gives $\bar{H}_{21,n} \tilde{\theta}_1 + \bar{H}_{22,n} \Delta_{\theta_{20}}^{\tilde{\theta}_2} = \check{H}_{22,n} \Delta_{\theta_{20}}^{\hat{\theta}_2}$. Adding and subtracting $\bar{H}_{22,n} \hat{\theta}_2$ to the LHS gives $\bar{H}_{21,n} \tilde{\theta}_1 + \bar{H}_{22,n} \Delta_{\theta_{20}}^{\tilde{\theta}_2} = \Delta_{\bar{H}_{22}}^{\check{H}_{22}} \Delta_{\theta_{20}}^{\hat{\theta}_2}$, solving which we get $\Delta_{\theta_{20}}^{\tilde{\theta}_2} = [\bar{H}_{22,n}^{-1} \bar{H}_{21,n}, \bar{H}_{22,n}^{-1} \Delta_{\bar{H}_{22}}^{\check{H}_{22}}] (\tilde{\theta}_1', \Delta_{\theta_{20}}^{\hat{\theta}_2})'$, implying

$$\Delta_{\theta}^{\hat{\theta}} = \mathcal{X}_n^{\bar{H}} \tilde{\theta}_1 + \left[0, \bar{H}_{22,n}^{-1} \Delta_{\bar{H}_{22}}^{\check{H}_{22}} \right] (0', \Delta_{\theta_{20}}^{\hat{\theta}_2})'. \quad (\text{A.19})$$

Substituting (A.19) into (A.15), it follows by techniques similar to those used in earlier proofs that \mathbf{LR}_n equals

$$\frac{n \tilde{\theta}_1' \mathcal{X}_n^{n^{-1} M'_n M} (n^{-1} M'_n M_n) \mathcal{X}_n^{m^{-1} M'_n M} \tilde{\theta}_1 - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} + o_p(1) = \frac{n \tilde{\theta}_1' (n^{-1} M'_n M_n^{11})^{-1} \tilde{\theta}_1 - q}{2^{\frac{1}{2}} q^{\frac{1}{2}}} + o_p(1),$$

by (3.4) and the partitioned matrix inverse formula. The proof now follows that for \mathbf{W}_n . \square

Proof of Theorem 3.3. Denote by $\mathbf{m}_{rt,n}$ the (r, t) -th element of \mathcal{M}_n . We seek to establish

asymptotic normality of

$$\frac{1}{2^{\frac{1}{2}}q^{\frac{1}{2}}} \sum_{r=1}^n (\epsilon_r^2 - 1) \mathbf{m}_{rr,n} + \frac{2}{2^{\frac{1}{2}}q^{\frac{1}{2}}} \sum_{r=1}^n \epsilon_r \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n} = \sum_{r=1}^n z_{rn}, \quad (\text{A.20})$$

say, with

$$z_{rn} = \frac{1}{2^{\frac{1}{2}}q^{\frac{1}{2}}} (\epsilon_r^2 - 1) \mathbf{m}_{rr,n} + \frac{2}{2^{\frac{1}{2}}q^{\frac{1}{2}}} \epsilon_r \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n}. \quad (\text{A.21})$$

We apply Theorem 2 of Scott (1973). Because the limit distribution is independent of the elements of M_n we can prove the result conditional on these and, henceforth, all expectations will be conditional on elements of M_n . First note that by Lemma B.1, $\sum_{r=1}^n z_{rn}$ has mean zero and variance converging to 1 in probability. Writing $\mathbb{1}(\cdot)$ for indicator function, we need to prove

$$\sum_{r=1}^n \mathbb{E} \{ z_{rn}^2 \mathbb{1}(z_{rn} \geq \epsilon) \} \xrightarrow{p} 0, \quad \forall \epsilon > 0, \quad (\text{A.22})$$

$$\sum_{r=1}^n \mathbb{E} (z_{rn}^2 \mid \epsilon_t, t < r) \xrightarrow{p} 1. \quad (\text{A.23})$$

To show (A.22) we can check the sufficient Lyapunov condition

$$\sum_{r=1}^n \mathbb{E} |z_{rn}|^{2+\frac{\chi}{2}} \xrightarrow{p} 0. \quad (\text{A.24})$$

Showing (A.24) boils down to proving that

$$\frac{1}{q^{1+\frac{\chi}{4}}} \sum_{r=1}^n \mathbb{E} \left(|\epsilon_r^2 - 1|^{2+\frac{\chi}{2}} \right) |\mathbf{m}_{rr,n}|^{2+\frac{\chi}{2}} \xrightarrow{p} 0, \quad (\text{A.25})$$

$$\frac{1}{q^{1+\frac{\chi}{4}}} \sum_{r=1}^n \mathbb{E} |\epsilon_r|^{2+\frac{\chi}{2}} \mathbb{E} \left| \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n} \right|^{2+\frac{\chi}{2}} \xrightarrow{p} 0 \quad (\text{A.26})$$

By Assumption 2, (A.25) is bounded by a constant times

$$\frac{n}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} |\mathbf{m}_{rr,n}|^{2+\frac{\chi}{2}} \quad (\text{A.27})$$

while by the Jensen, Burkholder, c_r , von Bahr-Esseen, triangle and Hölder inequalities

(A.26) is bounded by a constant times

$$\begin{aligned}
& \frac{n}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \mathbb{E} \left| \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n} \right|^{2+\frac{\chi}{2}} \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\mathbb{E} \left| \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n} \right|^{4+\chi} \right)^{\frac{1}{2}} \\
& \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\mathbb{E} \left| \sum_{t < r} \epsilon_t^2 \mathbf{m}_{rt,n}^2 \right|^{2+\frac{\chi}{2}} \right)^{\frac{1}{2}} \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\mathbb{E} \left| \sum_{t < r} \epsilon_t^4 \mathbf{m}_{rt,n}^4 \right|^{1+\frac{\chi}{4}} \right)^{\frac{1}{2}} \\
& \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\mathbb{E} \left| \sum_{t < r} (\epsilon_t^4 - \mu_4) \mathbf{m}_{rt,n}^4 \right|^{1+\frac{\chi}{4}} + \left| \sum_{t < r} \mathbf{m}_{rt,n}^4 \right|^{1+\frac{\chi}{4}} \right)^{\frac{1}{2}} \\
& \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\sum_{t < r} \mathbb{E} \left[|\epsilon_t^4 - \mu_4|^{1+\frac{\chi}{4}} \right] |\mathbf{m}_{rt,n}^4|^{1+\frac{\chi}{4}} + \left| \sum_{t < r} \mathbf{m}_{rt,n}^4 \right|^{1+\frac{\chi}{4}} \right)^{\frac{1}{2}} \\
& \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left\{ \left(\sum_{t < r} \mathbf{m}_{rt,n}^4 \right)^{1+\frac{\chi}{4}} \right\}^{\frac{1}{2}} \leq \frac{Cn}{q^{1+\frac{\chi}{4}}} \max_{1 \leq r \leq n} \left(\sum_{t < r} \mathbf{m}_{rt,n}^2 \right)^{1+\frac{\chi}{4}}. \tag{A.28}
\end{aligned}$$

For (A.27), note that $\mathbf{m}_{rt,n} = n^{-1} m'_{rn} \mathcal{X}_n^{n^{-1}M'M} (n^{-1} M'_n M_n)^{11} \mathcal{X}_n^{n^{-1}M'M} m_{tn}$ where m'_{rn} is the r -th row of M_n . Then

$$|\mathbf{m}_{rt,n}| = \mathcal{O}_p(n^{-1} \|m_{rn}\| \|m_{tn}\|) = \mathcal{O}_p(n^{-1} sm^2), \tag{A.29}$$

by Lemma B.6 and Assumption 3. This implies that (A.27) is $\mathcal{O}_p\left(\left(s^{\frac{1}{2}}m\right)^{4+\chi}/n^{1+\frac{\chi}{2}}q^{1+\frac{\chi}{4}}\right)$, which is negligible by (3.7). For (A.28), $\sum_{t=1}^n \mathbf{m}_{rt,n}^2$ equals n^{-1} times

$$\begin{aligned}
& m'_{rn} \mathcal{X}_n^{n^{-1}M'M} (n^{-1} M'_n M_n)^{11} \mathcal{X}_n^{n^{-1}M'M} (n^{-1} M'_n M_n) \mathcal{X}_n^{n^{-1}M'M} (n^{-1} M'_n M_n)^{11} \mathcal{X}_n^{n^{-1}M'M} m_{tn} \\
& \leq \|m_{rn}\|^2 \left\| \mathcal{X}_n^{n^{-1}M'M} (n^{-1} M'_n M_n)^{11} \mathcal{X}_n^{n^{-1}M'M} \right\|^2 \|n^{-1} M'_n M_n\| = \mathcal{O}_p(sm^2).
\end{aligned}$$

Then (A.28) is $\mathcal{O}_p\left(\left(s^{\frac{1}{2}}m\right)^{2+\frac{\chi}{2}}/n^{\frac{\chi}{4}}q^{1+\frac{\chi}{4}}\right)$, which is negligible by (3.7).

To show (A.23) note that by Lemma B.1 $\sum_{r=1}^n \mathbb{E}(z_{rn}^2 | \epsilon_t, t < r) - 1$ equals

$$\frac{1}{2q} \sum_{r=1}^n \sum_{\substack{t,u < r \\ t \neq u}} \epsilon_t \epsilon_u \mathbf{m}_{rt,n} \mathbf{m}_{ru,n} + \frac{1}{2q} \sum_{r=1}^n \sum_{t < r} (\epsilon_t^2 - 1) \mathbf{m}_{rt,n}^2 + \frac{1}{2q} \sum_{r=1}^n \mu_3 \mathbf{m}_{rr,n} \sum_{t < r} \epsilon_t \mathbf{m}_{rt,n}. \tag{A.30}$$

Also, because \mathcal{M}_n is idempotent and symmetric, we have, for $r, t = 1, \dots, n$,

$$\mathbf{m}_{rt,n} = \sum_{u=1}^n \mathbf{m}_{ru,n} \mathbf{m}_{ut,n} = \sum_{u=1}^n \mathbf{m}_{ru,n} \mathbf{m}_{tu,n}, \quad (\text{A.31})$$

i.e. \mathcal{M}_n and \mathcal{M}_n^2 have the same elements. In particular this indicates that

$$\text{tr}(\mathcal{M}_n^2) = \sum_r \sum_u \mathbf{m}_{ru,n} \mathbf{m}_{ru,n} = \sum_{r,u} \mathbf{m}_{ru,n}^2. \quad (\text{A.32})$$

The first term in (A.30) has zero mean and (conditional) variance bounded by $1/q^2$ times

$$\begin{aligned} C \sum_{\substack{v,r,t,u \\ t,u < r,v}} \mathbf{m}_{rt,n} \mathbf{m}_{ru,n} \mathbf{m}_{vt,n} \mathbf{m}_{vu,n} &= C \sum_{\substack{t,u \\ t,u < r,v}} \sum_r \mathbf{m}_{ru,n} \mathbf{m}_{rt,n} \sum_v \mathbf{m}_{vu,n} \mathbf{m}_{vt,n} \\ &= C \sum_{\substack{t,u \\ t,u < r,v}} \mathbf{m}_{tu,n}^2 \leq C \sum_{t,u} \mathbf{m}_{tu,n}^2 = Cq, \end{aligned} \quad (\text{A.33})$$

using (A.31), while the last equality follows because (A.32) and the idempotency of \mathcal{M}_n imply that $\sum_{r,t} \mathbf{m}_{rt,n}^2 = \text{tr}(\mathcal{M}_n^2) = \text{tr}(\mathcal{M}_n) = q$, because \mathcal{M}_n has rank q . Thus the first term in (A.30) is negligible because $q \rightarrow \infty$ as $n \rightarrow \infty$. The second term in (A.30) also has zero mean. By (A.29), (A.31) and (A.32), its variance is bounded by

$$\begin{aligned} Cq^{-2} \sum_{r,v} \sum_{t < r,v} \mathbf{m}_{rt,n}^2 \mathbf{m}_{vt,n}^2 &\leq Cq^{-2} \sum_{r,v,t} \mathbf{m}_{rt,n}^2 \mathbf{m}_{vt,n}^2 \\ &= Cq^{-2} \sum_{r,t} \mathbf{m}_{rt,n}^2 \sum_v \mathbf{m}_{vt,n}^2 = Cq^{-2} \sum_{r,t} \mathbf{m}_{rt,n}^2 \mathbf{m}_{tt,n} \\ &\leq Cq^{-2} \max_t |\mathbf{m}_{tt,n}| \sum_{r,t} \mathbf{m}_{rt,n}^2 = \mathcal{O}_p(q^{-1} n^{-1} s m^2), \end{aligned} \quad (\text{A.34})$$

which is again negligible under (3.7). Finally the third term in (A.30) has mean zero and variance bounded by $q^{-2} \max_t \mathbf{m}_{tt,n}^2 \sum_{r,t} \mathbf{m}_{rt,n}^2 = \mathcal{O}_p(q^{-1} n^{-2} s^2 m^4)$, which is negligible. \square

Proof of Theorem 3.4. First note that $n^{-1} M_n \left((n^{-1} M_n' M_n)^{-1} - \Sigma_n^{n^{-1} M' M} \right) M_n'$ is symmetric and has rank q , and when multiplied by itself gives

$$n^{-1} M_n \left\{ (n^{-1} M_n' M_n)^{-1} + n^{-1} \Sigma_n^{n^{-1} M' M} M_n' M_n \Sigma_n^{n^{-1} M' M} - 2 \Sigma_n^{n^{-1} M' M} \right\} M_n'.$$

Now partitioning $n^{-1} M_n' M_n$ yields $n^{-1} \Sigma_n^{n^{-1} M' M} M_n' M_n \Sigma_n^{n^{-1} M' M} = \Sigma_n^{n^{-1} M' M}$, so that \mathbf{W}_n

and \mathbf{LR}_n are quadratic forms in ϵ weighted by a symmetric, idempotent matrix with rank q . The proof now follows that of Theorem 3.3 and the details are omitted. \square

Proof of Theorem 3.5. (i): $\mathcal{C}.1$ follows from Theorems 3.1 and 3.3. For $\mathcal{C}.2$, $n\hat{g}'_{1,n}\hat{H}_n^{11}\hat{g}_{1,n} = ng_n'^0\mathcal{H}_ng_n^0$ by (A.2), where the 0 superscript denotes evaluation at $(0', \theta'_{20})'$, which is no longer the true parameter value under $H_{\ell 1}$. Another application of the MVT yields

$$g_n^0 = g_n - \frac{\partial}{\partial \theta'_1} g \left(\left(\begin{smallmatrix} \dot{\theta}'_1 \\ \theta'_{20} \end{smallmatrix} \right)' \right) \theta_{\ell,10} = g_n - \tau_n^{\dot{H}}, \quad (\text{A.35})$$

say, where $\|\dot{\theta}_1\| \leq \|\theta_{\ell,10}\|$. Thus

$$\frac{ng_n'^0\mathcal{H}_ng_n^0 - q}{2^{\frac{1}{2}}q^{\frac{1}{2}}} = \frac{ng_n'\mathcal{H}_ng_n - q}{2^{\frac{1}{2}}q^{\frac{1}{2}}} - 2\frac{n\tau_n'^{\dot{H}}\mathcal{H}_ng_n}{2^{\frac{1}{2}}q^{\frac{1}{2}}} + \frac{n\tau_n'^{\dot{H}}\mathcal{H}_n\tau_n'^{\dot{H}}}{2^{\frac{1}{2}}q^{\frac{1}{2}}}. \quad (\text{A.36})$$

The first term converges in distribution to a standard normal variate, as in Theorem 3.3, and is therefore $\mathcal{O}_p(1)$. Because at least one element of θ_{10} is nonzero, the third term on the RHS of (A.36) is readily seen by Assumption 2 and Lemma B.5 to be at least $\mathcal{O}_p\left(n/q^{\frac{1}{2}}\right)$, and evidently dominates the second term. Thus the LHS of (A.36) diverges at $n/q^{\frac{1}{2}}$ rate, and therefore so does \mathbf{LM}_n , by Theorem 3.1. Hence, for all $\varepsilon > 0$, $P\left(|\mathbf{LM}_n^{-1}| \leq \varepsilon q^{\frac{1}{2}}/n\right) \rightarrow 1$ as $n \rightarrow \infty$, implying that, for all $\varrho > 0$,

$$P(|\mathbf{LM}_n| > \varrho) \rightarrow 1, \text{ as } n \rightarrow \infty, \quad (\text{A.37})$$

because $q^{\frac{1}{2}} = o(n)$. Denote $\dot{\theta} = \left(\begin{smallmatrix} \dot{\theta}'_1 \\ \theta'_{20} \end{smallmatrix} \right)'$. To show $\mathcal{C}.3$ first note that under the sequence of local alternatives $H_{\ell 1}$, $\tau_n^{\dot{H}} = \left[\dot{H}_{11,n}, \dot{H}_{12,n} \right]' \theta_{\ell,10}$. As $\|\theta_{\ell,10}\| \rightarrow 0$, $\|\dot{\theta}_1\| \rightarrow 0$ also and so $\|\Delta_{\theta_0}^{\dot{\theta}}\| \rightarrow 0$ under $H_{\ell 1}$. Thus by Assumption 2 and (3.7), and because $\|\theta_{\ell,10}\|^2 = \mathcal{O}\left(q^{\frac{1}{2}}/n\right)$ we can use (A.36) to write

$$\frac{n\tau_n'^{\dot{H}}\mathcal{H}_n\tau_n^{\dot{H}}}{2^{\frac{1}{2}}q^{\frac{1}{2}}} = \mathcal{O}_p\left(\max\left\{\left\|\Delta_H^{\dot{H}}\right\|, \left\|\Delta_{n^{-1}M'M}^H\right\|, \left\|\Delta_L^{n^{-1}M'M}\right\|\right\}\right) + \frac{n\tau_n'^L\mathcal{L}_n\tau_n^L}{2^{\frac{1}{2}}q^{\frac{1}{2}}} \quad (\text{A.38})$$

$$\begin{aligned} \frac{n\tau_n'^{\dot{H}}\mathcal{H}_ng_n}{2^{\frac{1}{2}}q^{\frac{1}{2}}} &= \mathcal{O}_p\left(n^{\frac{1}{2}}q^{-\frac{1}{4}}\left\|\Delta_{n^{-1}M'\epsilon}^g\right\|\right) + \mathcal{O}_p\left(n^{-\frac{1}{2}}q^{-\frac{1}{4}}\|M'_n\epsilon\|\max\left\{\left\|\Delta_H^{\dot{H}}\right\|, \right. \right. \\ &\quad \left. \left. \left\|\Delta_{n^{-1}M'M}^H\right\|, \left\|\Delta_L^{n^{-1}M'M}\right\|\right\}\right) + \frac{\tau_n'^L\mathcal{L}_nM'_n\epsilon}{2^{\frac{1}{2}}q^{\frac{1}{2}}}, \end{aligned} \quad (\text{A.39})$$

with $\mathcal{L}_n = \mathcal{X}_n'^L L_n^{11} \mathcal{X}_n^L$. Everything apart from the last terms on the RHSs of (A.38) and (A.39) are negligible by (3.7). Choosing $\Gamma_n = [L_{11,n}, L_{12,n}] \mathcal{L}_n [L_{11,n}, L_{12,n}]'$ in $H_{\ell 1}$ the second term on the RHS of (A.38) equals $2^{-\frac{1}{2}}$, so that the LHS converges in probability to $2^{-\frac{1}{2}}$. The last term on the RHS of (A.39) has zero mean and conditional variance bounded by a constant times

$$\delta_n' [L_{11,n}, L_{12,n}] \mathcal{L}_n (n^{-1} M_n' M_n) \mathcal{L}_n [L_{11,n}, L_{12,n}]' \delta_n / q^{\frac{1}{2}} \delta_n' \Gamma_n \delta_n = \delta_n' [L_{11,n}, L_{12,n}] \mathcal{L}_n \times \\ \Delta_L^{n^{-1} M' M} \mathcal{L}_n [L_{11,n}, L_{12,n}]' \delta_n / q^{\frac{1}{2}} \delta_n' \Gamma_n \delta_n + 1/q^{\frac{1}{2}},$$

whose first term is $\mathcal{O}_p \left(\left\| \Delta_L^{n^{-1} M' M} \right\| / q^{\frac{1}{2}} \right) = o_p(1)$, by Assumption 3, using the techniques of Lemma B.5 to conclude that norms of blocks of L_n are $\mathcal{O}(1)$ and consequently so is \mathcal{L}_n . The simplification of the second term results from $\mathcal{L}_n L_n \mathcal{L}_n = \mathcal{L}_n$, because $\mathcal{X}_n'^L L_n \mathcal{X}_n^L = (L_n^{11})^{-1}$ by the partitioned inverse formula. Thus the LHS of (A.39) is negligible, proving the result.

Part (ii) is proved in an identical fashion and we omit the details, noting only that we take $\Gamma_n = [L_{11,n}, L_{12,n}] (L_n^{-1} - \Sigma_n^L) [L_{11,n}, L_{12,n}]'$. \square

Proof of Proposition 4.1. The condition (4.12) is simply the union of those imposed in Theorem 3.5 and by Newey (1997) to obtain $\sup_{x \in \mathcal{X}} \left| \hat{d}(x) - d(x) \right| = o_p(1)$ and asymptotic normality. We have $\|n^{-1} P' P - \mathbb{E}(P_J(x_i) P_J(x_i)')\| = \mathcal{O}_p \left(\zeta(J) J^{\frac{1}{2}} / n^{\frac{1}{2}} \right)$ and $\|P' v\| = \mathcal{O}_p \left((Jn)^{\frac{1}{2}} \right)$ (due to x_i iid and (4.11)). Together with (4.6) these imply that (3.7) holds under (4.12). \square

Assumption A.* (Robinson (2005, 2010)) Define $\varphi = (1 + \phi(t_1)) (\phi(t_1) - \phi(t_2))^{-1}$ where $[t_1, t_2]$ is an interval on which $f(t)$ is bounded away from zero, and $\varkappa = 1 + 2^{\frac{1}{2}} \simeq 2.414$, The following conditions hold:

- The v_i are iid with zero mean, unit variance and differentiable pdf $f(t)$.
- $\phi_\ell(t) = \phi_t^\ell$ with $\phi(t)$ strictly increasing and thrice differentiable such that for some $\kappa \geq 0$ and $K < \infty$, $|\phi(t)| \leq 1 + |t|^\kappa$ and $|\phi_t(t)| + |\phi_{tt}(t)| + |\phi_{ttt}(t)| \leq C \left(1 + |\phi(t)|^K \right)$.
- $J \rightarrow \infty$ as $n \rightarrow \infty$, and either (i) $\kappa = 0$, $\mathbb{E}(v_i^4) < \infty$, and

$$\lim_{n \rightarrow \infty} (\log n / J) > 8 \{ \log \varkappa + \max(\log \varphi, 0) \} \simeq 7.05 + 8 \max(\log \varphi, 0);$$

or (ii) $\kappa > 0$, for some $\omega > 0$ the moment generating function $\mathbb{E}(\exp(u |v_i|^\omega))$ exists

for some $u > 0$, and

$$\lim_{n \rightarrow \infty} (\log n / J \log J) \geq \max \{8\kappa\omega^{-1}, 4\kappa(\omega + 1)\omega^{-1}\};$$

or (iii) $\kappa > 0$, v_i is almost surely bounded, and

$$\lim_{n \rightarrow \infty} (\log n / J \log J) \geq 4\kappa.$$

- Define ρ_{vJ} as CJ if $v = 0$, $(CJ)^{vJ/\omega}$ if $v > 0$ and Assumption A* (ii) holds and C^J , if $v > 0$ and Assumption A* (iii) holds. Then $J^{\frac{1}{2}}\rho_{4\kappa J}^{\frac{1}{2}}/n^{\frac{1}{2}}(\rho_{2\kappa J}/t(J) + 1) \rightarrow 0$ as $n \rightarrow \infty$.

Proof of Proposition 5.1. The proof follows straightforwardly because as in Newey (1988) we can use $e_i = y_n - X_n\hat{\beta}_{OLS}$ and prove the result with e_i replaced by v_i . Then $\mathbb{E} \|\bar{W}^{(J)}u\|^2 = \sum_{i=1}^n \mathbb{E} \|\Phi^{(J)}(v_i)\|^2 = \mathcal{O}_p(n\rho_{2\kappa J})$ by (A.37) of Robinson (2010) (note that we do not have the t_{iin} factor that arises in his SAR setup), whence $\|\bar{W}^{(J)}u\| = \mathcal{O}_p\left(n^{\frac{1}{2}}\rho_{2\kappa J}^{\frac{1}{2}}\right)$ by Markov's inequality. Similarly by Lemma 10 of Robinson (2005) $\|W^{(J)} - \mathbb{E}\{\Phi^{(J)}(v_i)\Phi'^{(J)}(v_i)\}\| = \mathcal{O}_p\left(J^{\frac{1}{2}}\rho_{4\kappa J}^{\frac{1}{2}}/n^{\frac{1}{2}}\right)$. Now the conditions of the Proposition ensure that (3.4) and (3.7) are satisfied. \square

B Technical lemmas

Lemma B.1. *Let the conditions of Theorem 3.3 hold. Conditional on elements of M_n , $\mathbb{E}(\epsilon' \mathcal{M}_n \epsilon - q) = 0$ and $\text{var}(\epsilon' \mathcal{M}_n \epsilon) / 2q \xrightarrow{p} 1$.*

Proof. The first part follows from $\mathbb{E}(\epsilon' \mathcal{M}_n \epsilon) = \text{tr} \mathcal{M}_n = q$, by symmetry, idempotency and rank q of \mathcal{M}_n . These properties also imply $\text{var}(\epsilon' \mathcal{M}_n \epsilon) = (\mu_4 - 3) \sum_r \mathbf{m}_{rr,n}^2 + \text{tr}(\mathcal{M}_n \mathcal{M}_n') + \text{tr}(\mathcal{M}_n^2) = \mathcal{O}_p(n^{-1}s^2m^4) + 2q$, by (A.29), so $\text{var}(\epsilon' \mathcal{M}_n \epsilon) / 2q = \mathcal{O}_p(q^{-1}n^{-1}s^2m^4) + 1 = o_p(1) + 1$, by (3.7). \square

Lemma B.2. *Under the conditions of Theorem 3.1 or Theorem 3.2,*

$$\|\hat{H}_n^{-1}\| = \mathcal{O}_p(\|H_n^{-1}\|) = \mathcal{O}_p\left(\|(n^{-1}M_n' M_n)^{-1}\|\right) = \mathcal{O}_p\left(\left\{\lim_{n \rightarrow \infty} \underline{\eta}(L_n)\right\}^{-1}\right) = \mathcal{O}_p(1).$$

A similar result holds for evaluation at $\tilde{\theta}$ and $\bar{\theta}$ satisfying $\|\Delta_{\theta_0}^{\bar{\theta}}\| \leq \|\Delta_{\theta_0}^{\hat{\theta}}\|$ or $\|\Delta_{\theta_0}^{\bar{\theta}}\| \leq$

$\left\| \Delta_{\theta_0}^{\tilde{\theta}} \right\|$. The result also implies that the norms of blocks of partitioned versions of these matrices are bounded.

Proof. See Lemma B.8 in Gupta and Robinson (2015b). For square blocks, the last claim follows by Lemma B.4. For non-square blocks, it follows as in the proof of Lemma B.5 (see (B.2)). \square

Lemma B.3. *Under the conditions of Theorem 3.1 or Theorem 3.2,*

$$\left\| \hat{H}_n \right\| = \mathcal{O}_p(\|H_n\|) = \mathcal{O}_p(\|n^{-1}M'_n M_n\|) = \mathcal{O}_p\left(\overline{\lim}_{n \rightarrow \infty} \bar{\eta}(L_n)\right) = \mathcal{O}_p(1).$$

A similar result holds for evaluation at $\tilde{\theta}$ and $\bar{\theta}$ satisfying $\left\| \Delta_{\theta_0}^{\bar{\theta}} \right\| \leq \left\| \Delta_{\theta_0}^{\hat{\theta}} \right\|$ or $\left\| \Delta_{\theta_0}^{\bar{\theta}} \right\| \leq \left\| \Delta_{\theta_0}^{\tilde{\theta}} \right\|$. The result also implies that the norms of blocks of partitioned versions of these matrices are bounded.

Proof. Similar to proof of Lemma B.2. \square

Lemma B.4. *Consider a square matrix A_n . Suppose that $\{\underline{\eta}(A_n)\}^{-1} + \bar{\eta}(A_n) = \mathcal{O}_p(1)$. Let A_n^S be any diagonal square sub-matrix of A_n . Then $\{\underline{\eta}(A_n^S)\}^{-1} + \bar{\eta}(A_n^S) = \mathcal{O}_p(1)$.*

Proof. Follows easily by appropriate choice of x_n in quadratic form $x'_n A_n x_n$. \square

Lemma B.5. *Under the conditions of Theorem 3.1 or Theorem 3.2, $\|\mathcal{H}_n\| = \mathcal{O}_p(1)$.*

Proof. By elementary norm inequalities and properties

$$\|\mathcal{H}_n\| \leq \left\| \mathcal{X}_n^{\bar{H}} \right\|^2 \left\| \hat{H}_n^{11} \right\|. \quad (\text{B.1})$$

For the first factor on the RHS of (B.1), note that $\left\| \mathcal{X}_n^{\bar{H}} \right\| \leq \|I_q\| + \left\| \bar{H}_{22,n}^{-1} \right\| \left\| \bar{H}_{21,n} \right\|$. Also $\bar{H}_{21,n} = I_{1,n} \bar{H}_n I_{2,n}$ where $I_{1,n} = [0_{(s-q) \times q}, I_{s-q}]$ and $I_{2,n} = [I_q, 0_{q \times (s-q)}]'$. Clearly $\|I_{1,n}\|, \|I_{2,n}\| \leq 1$. Then

$$\left\| \bar{H}_{21,n} \right\| \leq \|I_{1,n}\| \|I_{2,n}\| \left\| \bar{H}_n \right\| \leq \left\| \bar{H}_n \right\| = \mathcal{O}_p(1), \quad (\text{B.2})$$

by Lemma B.3. By the partitioned inverse formula, we have $\bar{H}_{22,n}^{-1} = \bar{H}_n^{22} - \bar{H}_n^{21} \left(\bar{H}_n^{22} \right)^{-1} \bar{H}_n^{12}$ implying that $\left\| \bar{H}_{22,n}^{-1} \right\| \leq \left\| \bar{H}_n^{22} \right\| + \left\| \bar{H}_n^{21} \right\|^2 \left\| \left(\bar{H}_n^{22} \right)^{-1} \right\|$. Since eigenvalues of diagonal sub-matrices fall within the same bounds as those of the original matrix (see Lemma B.4)

$\left\| \left(\overline{H}_n^{22} \right)^{-1} \right\| = \left\{ \underline{\eta} \left(\overline{H}_n^{22} \right) \right\}^{-1} = \mathcal{O}_p(1)$. Arguments similar to those used for showing (B.2) allow us to conclude that $\left\| \overline{H}_n^{22} \right\|$ and $\left\| \overline{H}_n^{21} \right\|$ are $\mathcal{O}_p(1)$ using Lemma B.3. It follows that the first factor on the RHS of (B.1) is $\mathcal{O}_p(1)$. The second factor is also $\mathcal{O}_p(1)$ by Lemmas B.2 and B.4. \square

Lemma B.6. *Under the conditions of Theorem 3.1 or Theorem 3.2,*

$$\left\| \mathcal{X}_n^{n^{-1}M'M} \right\| + \|n^{-1}M'_n M_n\| = \mathcal{O}_p(1).$$

Proof. The LHS is bounded by $\left\| \Delta_{\mathcal{X}^L}^{\mathcal{X}_n^{n^{-1}M'M}} \right\| + \left\| \Delta_L^{n^{-1}M'M} \right\| + \|\mathcal{X}_n^L\| + \|L_n\|$. The last two terms are $\mathcal{O}(1)$ by Assumption 2 and the techniques used in the proof of Lemma B.5, while the first is $\left\| \Delta_{L_{12}L_{22}^{-1}}^{n^{-1}M'M_{12}n^{-1}M'M_{22}^{-1}} \right\| = \mathcal{O}_p \left(\left\| \Delta_L^{n^{-1}M'M} \right\| \right)$, as we bounded $\left\| \Delta_{\overline{H}_{12}\overline{H}_{22}^{-1}}^{L_{12}L_{22}^{-1}} \right\|$ in the proof of Theorem 3.1. As $\left\| \Delta_L^{n^{-1}M'M} \right\| = o_p(1)$ by (3.4), the lemma is proved. \square

References

- Ackerberg, D., X. Chen, and J. Hahn (2012). A practical asymptotic variance estimator for two-step semiparametric estimators. *Review of Economics and Statistics* 94, 481–498.
- Ahmad, I., S. Leelahanon, and Q. Li (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *Annals of Statistics* 33, 258–283.
- Andrews, D. W. K. (1985). A zero-one result for the least squares estimator. *Econometric Theory* 1, 85–96.
- Andrews, D. W. K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–345.
- Andrews, D. W. K. (1994). Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica* 62, 43–72.
- Beran, R. (1976). Adaptive estimates for autoregressive processes. *Annals of the Institute of Statistical Mathematics* 28, 77–89.
- Berk, K. N. (1974). Consistent autoregressive spectral estimates. *Annals of Statistics* 2, 489–502.
- Case, A. C. (1991). Spatial patterns in household demand. *Econometrica* 59, 953–965.
- Case, A. C. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics* 22, 491–508.
- Cliff, A. D. and J. K. Ord (1973). *Spatial Autocorrelation*. London: Pion.
- de Jong, R. M. and H. J. Bierens (1994). On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity. *Econometric Theory* 10(01), 70–90.
- Eubank, R. L. and C. H. Spiegelman (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *Journal of the American Statistical Association* 85, 387–392.
- Fan, Y. and Q. Li (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* 64, 865–890.

- Gallant, R. and D. W. Nychka (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica* 55, 363–390.
- Gupta, A. and P. M. Robinson (2015a). Inference on higher-order spatial autoregressive models with increasingly many parameters. *Journal of Econometrics* 186, 19–31.
- Gupta, A. and P. M. Robinson (2015b). Pseudo maximum likelihood estimation of higher-order spatial autoregressive models with increasing dimension. Mimeo.
- Gurmu, S., P. Rilstone, and S. Stern (1999). Semiparametric estimation of count regression models. *Journal of Econometrics* 88, 123–150.
- Hong, Y. and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica* 63, 1133–1159.
- Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics* 1, 799–821.
- Jenish, N. (2014). Spatial semiparametric model with endogenous regressors. *Econometric Theory FirstView*, 1–26.
- Kelejian, H. H. and I. R. Prucha (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics* 17, 99–121.
- Lavergne, P. and Q. Vuong (2000). Nonparametric significance testing. *Econometric Theory* 16, 576–601.
- Lee, J. and P. M. Robinson (2013). Series estimation under cross-sectional dependence. Preprint, London School of Economics.
- Lee, L. F. (2002). Consistency and efficiency of least squares estimation for mixed regressive, spatial autoregressive models. *Econometric Theory* 18, 252–277.
- Lee, L. F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72, 1899–1925.
- Lorentz, G. G. (1986). *Approximation of Functions*. Chelsea.
- Newey, W. K. (1988). Adaptive estimation of regression models via moment restrictions. *Journal of Econometrics* 38, 301–339.

- Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* 79, 147–168.
- Portnoy, S. (1984). Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large. I. Consistency. *Annals of Statistics* 12, 1298–1309.
- Portnoy, S. (1985). Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large; II. Normal approximation. *Annals of Statistics* 13, 1403–1417.
- Robinson, P. M. (1979). Distributed lag approximation to linear time-invariant systems. *Annals of Statistics* 7, 507–515.
- Robinson, P. M. (1988). Root- n -consistent semiparametric regression. *Econometrica* 56, 931–954.
- Robinson, P. M. (2003). Denis Sargan: Some perspectives. *Econometric Theory* 19, 481–494.
- Robinson, P. M. (2005). Efficiency improvements in inference on stationary and nonstationary fractional time series. *Annals of Statistics* 33, 1800–1842.
- Robinson, P. M. (2010). Efficient estimation of the semiparametric spatial autoregressive model. *Journal of Econometrics* 157, 6–17.
- Scott, D. J. (1973). Central limit theorems for martingales and for processes with stationary increments using a Skorokhod representation approach. *Advances in Applied Probability* 5, 119–137.
- Su, L. and S. Jin (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics* 157, 18–33.
- Wooldridge, J. M. (1992). A test for functional form against nonparametric alternatives. *Econometric Theory* 8, 452–475.
- Yatchew, A. (2000). Scale economies in electricity distribution: a semiparametric analysis. *Journal of Applied Econometrics* 210, 187–210.
- Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

- Yatchew, A. J. (1992). Nonparametric regression tests based on least squares. *Econometric Theory* 8, 435–451.
- Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75, 263 – 289.